Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

Ανάλυση και Εφαρμογές Μηχανικής Μάθησης σε Δεδομένα Εκπαιδευτικών Οικοσυστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γκουβάτσου Χριστίνα Αικατερίνη

Επιβλέπων: Πέτρος Στεφανέας
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2023

Στη μητέρα μου, τον Κωνσταντίνο, τη Μαντώ, την Πρίσκιλλα και την κα Γρηγορίου

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

Ανάλυση και Εφαρμογές Μηχανικής Μάθησης σε Δεδομένα Εκπαιδευτικών Οικοσυστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γκουβάτσου Χριστίνα Αικατερίνη

Επιβλέπων: Πέτρος Στεφανέας
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Νοεμβρίου 2023.

.....	.....	.....
Πέτρος Στεφανέας	Βασίλειος Μάγκλαρης	Μαίρη Γραμματικού
Αν. Καθηγητής Ε.Μ.Π.	Ομότιμος Καθηγητής Ε.Μ.Π.	ΕΔΙΠ ΣΗΜΜΥ/ Ε.Μ.Π.

Αθήνα, Νοέμβριος 2023

.....
Γκουβάτσου Χριστίνα Αικατερίνη
Διπλωματούχος Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών Ε.Μ.Π.

Copyright ©Γκουβάτσου Χριστίνας Αικατερίνης, 2023.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάλυση και μελέτη δεδομένων Web Analytics που προκύπτουν από την χρήση του εκπαιδευτικού οικοσυστήματος Up2University (Up2U). Το εκπαιδευτικό αυτό ψηφιακό οικοσύστημα (Virtual Learning Environment) δημιουργήθηκε στα πλαίσια ενός ευρωπαϊκού ερευνητικού προγράμματος με στόχο την αποτελεσματικότερη προετοιμασία των μαθητών για την τριτοβάθμια εκπαίδευση, και η δομή του βασίστηκε στις αρχές του Next Generation Digital Learning Environment (NGDLE).

Τα δεδομένα αφορούν το στιγμιότυπο (e-Service) που δημιουργήθηκε για σχολεία της Πολωνίας που συμμετείχαν στο πρόγραμμα και συλλέχθηκαν από το ανοιχτό λογισμικό συλλογής Web Analytics το Matomo κατά την περίοδο 8 Νοεμβρίου 2019 έως 25 Μαρτίου 2021 σε ακατέργαστη μορφή. Επομένως, ήταν απαραίτητη η πλήρης κατανόηση των μεταβλητών και των τιμών τους, ο ενδελεχής καθαρισμός τους από θόρυβο, η αποκατάσταση των ελλιπών τιμών και η δημιουργία κατάλληλων πινάκων. Για την αντιμετώπιση των ελλιπών τιμών έγινε χρήση της Επιβλεπόμενης Μεθόδου Μηχανικής Μάθησης k-Πλησιέστερων Γειτόνων (kNN).

Η ανάλυση πραγματοποιήθηκε σε όλο το σύνολο δεδομένων, με το ενδιαφέρον της περιγραφικής και διευρευνητικής ανάλυσης να επικεντρώνεται στη χρήση της πλατφόρμας κατά τη διάρκεια των κυμάτων της πανδημίας Covid-19 και της υποχρεωτικής τηλεκπαίδευσης, και στο το πώς αυτό επηρέασε τις μετρικές που αφορούν τους επισκέπτες (Visitor Type), τη δραστηριότητα (Activity Metrics) και τη μετρική Bounce Rate(%). Στη συνέχεια, για τη μετρική Bounce Rate(%) εφαρμόζονται τα μοντέλα ταξινόμησης Logistic Regression, Random Forest, Gradient Boosting, XGBoost, SVM ώστε να αξιολογηθούν και να συγκριθούν ως προς την ικανότητά τους να κάνουν προβλέψεις και να ταξινομούν τη μετρική Bounce Rate(%) σε Low Bounce Rate και High Bounce Rate. Τα μοντέλα εκπαιδεύονται πρώτα με όλες τις μεταβλητές στο σύνολο δεδομένων σε baseline μορφή, εν συνεχεία με εφαρμογή της μεθόδου διαδοχικής απαλοιφής μεταβλητών Recursive Feature Elimination (RFE), και τέλος ερμηνεύεται η σημαντικότητα των μεταβλητών για τις δύο περιπτώσεις για την καλύτερη κατανόηση των μοντέλων.

Λέξεις κλειδιά: Up2University (Up2U), Covid-19, Web Analytics, Next Generation Digital Learning Environment (NGDLE), VLE, Matomo, Supervised Machine Learning, Exploratory Data Analysis, kNN Αλγόριθμος, Ταξινόμηση, Random Forest, Gradient Boosting, XGBoost, SVM, Feature Importance

Abstract

The purpose of this thesis is to perform an analysis and apply machine learning models on web analytics extracted from the virtual learning environment (VLE) of the Up2University (Up2U) platform. The VLE was created as part of a European research project aiming to better prepare students for the Tertiary Education, and it was based on the principles of Next Digital Learning Environment (NGDLE).

The data were collected in raw form, from the open-source web analytics tracker Matomo, correspond to the polish e-service created for polish schools participating in the project. The data entries start from 8th November 2019 and end on the 25th of March 2021. Exploratory data analysis was performed in order to understand the features, their relationships, identify patterns and clean the data from missing values using kNN Imputation.

The analysis was conducted on the whole range, however the descriptive part was segmented into periods based on the Covid-19 pandemic waves in Poland and its impact was studied on the metrics regarding Visitor Types, Activity Metrics, and the Bounce Rate(%), for these periods. Next, machine learning classifiers are applied such as Logistic Regression, Random Forest, Gradient Boosting, XGBoost and SVM are trained in their baseline form on the whole data set and using all features. Their performance is evaluated on their ability to predict the classes of the metric Bounce Rate(%) in classes of Low Bounce Rate(%) and High Bounce Rate(%). After that, Recursive Feature Elimination is performed and the models are re-trained and re-evaluated. Finally, the feature importance is calculated and discussed for each model, in an attempt to further understand and explain the models' behaviour.

Keywords: Up2University (Up2U), Covid-19, Web Analytics, Next Generation Digital Learning Environment (NGDLE), VLE, Matomo, Supervised Machine Learning, Exploratory Data Analysis, kNN Imputation, Classification, Random Forest, Gradient Boosting, XGBoost, SVM, Feature Importance

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους δύο καθηγητές μου, τον κ. Βασίλειο Μάγκλαρη και τον κ. Πέτρο Στεφανέα, για την τιμή να επιβλέπουν την πορεία μου σε αυτή την διπλωματική, τον υποψήφιο διδάκτορα της Σχολής ΗΜΜΥ Δημήτρη Πανταζάτο και την ερευνήτρια της Σχολής ΗΜΜΥ κα Μαίρη Γραμματικού για τον χρόνο και τις οδηγίες που διέθεσαν.

Ένα μεγάλο ευχαριστώ οφείλω στην οικογένειά μου, στα ξαδέρφια μου Γιάννη και Δημήτρη, για την αμέριστη υποστήριξή τους σε κάθε επιλογή μου και προπάντων στη μητέρα μου για την αγάπη της, τις αξίες της και το ήθος της. Τη διπλωματική αυτή την αφιερώνω στον αδελφό μου Κωνσταντίνο, τις αδελφικές μου φίλες Μαντώ και Πρίσκιλλα για την πίστη τους σε εμένα, τη συνεχή ενθάρρυνση και την αγάπη τους που με οδηγεί στο να γίνομαι η καλύτερη εκδοχή του εαυτού μου καθώς και στην κα Γρηγορίου για όλους τους ελεύθερους συνειρμούς μου.

Χριστίνα

Περιεχόμενα

1 Up2University Project **2**
- 1.1 Στόχοι Προγράμματος . . . 2
- 1.2 Υπηρεσίες του Up2U . . . 3
- 1.3 Αρχές του Up2U . . . 4
- 1.4 Αρχιτεκτονική και Σχεδιασμός του Up2U . . . 6
  - 1.4.1 Στιγμιότυπα - e-Service . . . 7

2 Web Analytics **11**
- 2.1 Web Analytics . . . 11
  - 2.1.1 Πώς χρησιμοποιούνται τα Web Analytics . . . 13
  - 2.1.2 Μετρικές Web Analytics . . . 15

3 Θεωρητικό Υπόβραθρο **19**
- 3.1 Καθαρισμός & Προεπεξεργασία Δεδομένων . . . 19
  - 3.1.1 Χειρισμός Ελλιπών Τιμών . . . 19
  - 3.1.2 k-Nearest Neighbors Imputation . . . 21
- 3.2 Μοντέλα Μηχανικής Μάθησης . . . 23
  - 3.2.1 Μοντέλα Ταξινόμησης . . . 24
    - 3.2.1.1 Logistic Regression . . . 24
    - 3.2.1.2 Random Forest . . . 25
    - 3.2.1.3 Gradient Boosting και XGBoost . . . 27
    - 3.2.1.4 Support Vector Machine . . . 28
  - 3.2.2 Μετρικές και Μέθοδοι Αξιολόγησης Μοντέλων Ταξινόμησης . . . 30
    - 3.2.2.1 Cross Validation . . . 30
    - 3.2.2.2 Confusion Matrix . . . 31
    - 3.2.2.3 Accuracy . . . 32
    - 3.2.2.4 Precision, Recall & F1-score . . . 32
  - 3.2.3 Feature Selection και Feature Importance . . . 33

4 Επεξεργασία και καθαρισμός δεδομένων **36**
- 4.1 Μετατροπή Τύπων Μεταβλητών . . . 36
- 4.2 Εκτίμηση Ελλιπών Παρατηρήσεων με kNN Imputation . . . 41
- 4.3 Επιλογή Χρονικών Περιόδων . . . 46

5 Περιγραφική Ανάλυση και Εφαρμογή Μοντέλων Μηχανικής Μάθησης **48**

5.1 Περιγραφική & Διερευνητική Ανάλυση 48
5.1.1 Περίοδος Pre Covid-19 51
5.1.2 Περίοδος Wave 1 57
5.1.3 Περίοδος Wave 2 62
5.1.4 Περίοδος Post Covid-19 67
5.1.5 Συσχετίσεις & Πολυσυγγραμμικότητα στο σύνολο δεδομένων polish 72
5.1.6 Σύνοψη Αποτελεσμάτων 74
5.2 Εφαρμογή Μοντέλων Μηχανικής Μάθησης 75
5.2.1 Εφαρμογή Μοντέλων - Baseline 78
5.2.1.1 Σύνοψη αποτελεσμάτων και σύγκριση μεθόδων 88
5.2.2 Εφαρμογή με Recursive Feature Elimination 89
5.2.2.1 Συμπεράσματα 95
5.2.3 Feature Importance και Feature Ranking 96
5.2.3.1 Feature Importance Baseline 96
5.2.3.2 Feature Importance με RFE 100

6 Επίλογος **106**
6.1 Σύνοψη 106
6.2 Προοπτικές Μελλοντικής Επεκτάσεων 107

Βιβλιογραφία **108**

Κεφάλαιο 1

Up2University Project

1.1 Στόχοι Προγράμματος

Το Up To University, ή Up2U εν συντομία, είναι ένα ηλεκτρονικό οικοσύστημα, βασισμένο στις προδιαγραφές του Next Generation Digital Learning Environment (NGDLE) που δημιουργήθηκε στα πλαίσια του ευρωπαϊκού προγράμματος "Up to University - Bridging the gap between schools and universities through informal education". Το πρόγραμμα ξεκίνησε 1 Ιανουαρίου 2017 και ολοκληρώθηκε στις 31 Μαΐου 2020, μετρώντας τη συνεργασία του Πανευρωπαϊκού Ακαδημαϊκού-Ερευνητικού Δικτύου GÉANT, 18 μελών από 12 ευρωπαϊκές χώρες, συμπεριλαμβανομένων Εθνικών Δικτύων Υποδομών Τεχνολογίας και Έρευνας, πανεπιστημίων και παρόχους υποδομών.

Κύριος στόχος του Up2U προγράμματος είναι η γεφύρωση του χάσματος μεταξύ δευτεροβάθμιας και τριτοβάθμιας εκπαίδευσης, με την ενσωμάτωση σύγχρονων τεχνολογικών και ψηφιακών εκπαιδευτικών υπηρεσιών. Το πρόγραμμα στηρίχθηκε, επίσης, στην αξιοποίηση τυπικών (Formal) και μη-τυπικών (Non Formal) μεθόδων εκπαίδευσης. Συνεπώς, η συνδυασμένη εφαρμογή συγχρόνων και ασύγχρονων πρακτικών σε αυτό το διαδραστικό περιβάλλον βάσει Project Based Learning (PBL) μεθόδων και την αμοιβαία υποστήριξη μεταξύ των μαθητών παρέχει μια πιο ολοκληρωμένη και δυναμική εκπαιδευτική εμπειρία που προετοιμάζει τους μαθητές για τις προκλήσεις του πανεπιστημίου.

1.2 Υπηρεσίες του Up2U

Οι υπηρεσίες του Up2U βασίζονται στην ενσωμάτωση και συνδυασμό διαφόρων εργαλείων σε ένα έξυπνο και καινοτόμο οικοσύστημα. Παράλληλα, αυτοματοποιημένες προσθήκες (add-ons) και ρυθμίσεις για κάθε εργαλείο καθιστούν την πλατφόρμα ακόμη πιο ισχυρή. Οι υπηρεσίες του Up2U είναι δυνατό να χρησιμοποιηθούν είτε μεμονωμένα, είτε ως μέρος ολόκληρου του Up2U οικοσυστήματος.

Ένα σημαντικό πλεονέκτημα αυτού του οικοσυστήματος είναι ότι οι χρήστες έχουν, ασφαλή, πρόσβαση σε οποιαδήποτε από τις υπηρεσίες του μέσω ενός μοναδικού λογαριασμού, χωρίς να απαιτείται δημιουργία διαφορετικών λογαριασμών για κάθε εργαλείο. Αυτή η Single Sign On (SSO) ταυτοποίηση στοιχείων, διασφαλίζει την ασφάλεια των προσωπικών στοιχείων του εκάστοτε χρήστη.

Οι υπηρεσίες του Up2U που αποτελούν το Up2U Application Toolbox είναι οι εξής:

- **Moodle** - μία Learning Management System(LMS) πλατφόρμα βασισμένη σε λογισμό ανοιχτού κώδικα (open source software) που βοηθά στην οργάνωση και επίβλεψη των διαδικτυακών μαθημάτων
- **eduMEET** - πλατφόρμα ανοιχτού λογισμικού, με ασφαλή λειτουργία στο διαδίκτυο για βιντεο-διασκέψεις εκπαιδευτικού χαρακτήρα.
- **CERNBox** - υπηρεσία που διατίθεται μέσω υπολογιστικού νέφους (cloud-based) και έχει σχεδιαστεί στο CERN. Εξυπηρετεί στην αποθήκευση, στο συγχρονισμό και στην κοινοποίηση αρχείων μεταξύ χρηστών ή ομάδων χρηστών. Η είσοδος στην υπηρεσία είναι εφικτή μέσω υπολογιστή, κινητού τηλεφώνου, ή του διαδικτύου.
- **SWAN** - Service for Web-based Analysis (SWAN), μία πλατφόρμα που επιτρέπει τη δημιουργία διαδραστικών, ψηφιακών τετραδίων (notebook). Αυτά τα τετράδια μπορούν να περιέχουν εικόνες, βίντεο, προγραμματιστικό κώδικα και εξισώσεις. Η πρόσβασή τους είναι εύκολη και πραγματοποιείται μέσω του CERNBox.
- **eduOER** - υπηρεσία που δίνει πρόσβαση σε άλλες, δωρεάν εκπαιδευτικές πηγές για καλύτερη ανταλλαγή πληροφοριών.
- **SelCont** - Synchronised eLearning Content (SelCont), η υπηρεσία που συγχρονίζει τα βίντεο των διαλέξεων με τις παρουσιάσεις του μαθήματος. Απαιτεί ελάχιστη επεξεργασία για την προετοιμασία του μαθήματος, ενώ δεν προϋποθέτει ιδιαίτερα πολλές ρυθμίσεις στην αίθουσα διαλέξεων. Η πρόσβαση είναι δυνατή από υπολογιστή, τάμπλετ, ή κινητό τηλέφωνο. Σχεδιάστηκε από το εργαστήριο Network Management & Optimal Design (NETMODE) του ΕΜΠ.
- **DSpace** - αποθηκευτικός χώρος στο υπολογιστικό νέφος που επιτρέπει τους χρήστες, είτε εκπαιδευτικούς είτε μαθητές, να αποθηκεύουν αρχεία, βίντεο, φωτογραφίες, αρχεία κ.λ.π.. Η πρόσβαση είναι εφικτή απευθείας από το DSpace, ή από την υπηρεσία Moodle του Up2U.

1.3 Αρχές του Up2U

Πέραν της τεχνολογικής υποδομής του Up2U και των προαναφερθέντων τεχνικών υπηρεσιών, το Up2U πρόγραμμα σχεδιάστηκε και βάσει παιδαγωγικών θεωριών καθώς και της ιδέας πως οι μαθητές πρέπει να διαθέτουν συγκεκριμένες δεξιότητες και εφόδια για την ακαδημαϊκή και επαγγελματική τους σταδιοδρομία στον 21ο αιώνα. Μέσω αυτού του συνδυασμού της τεχνολογικής και παιδαγωγικής σκοπιάς, το Up2U πρόγραμμα στοχεύει στην ανάπτυξη των ικανοτήτων των μαθητών και την προετοιμασία τους από το σχολείο.

Παρακάτω παραθέτονται οι Αρχές – επιθυμητές δεξιότητες – του Up2U με μια μικρή περιγραφή για την κάθε αρχή.

- **Information** (Αρχή Πληροφορίας):

  Αναγνώριση, εντοπισμός, επανάκτηση, αποθήκευση, οργάνωση και ανάλυση ψηφιακών πληροφοριών, ανάλογα με τη σχετικότητά τους και τον σκοπό που εξυπηρετούν.

- **Communication** (Αρχή Επικοινωνίας):

  Επικοινωνία σε ψηφιακό επίπεδο και περιβάλλον, κοινοποίηση πηγών μέσω ηλεκτρονικών εργαλείων. Ενθαρρύνεται ακόμη η διάδραση και συνεργασία με άλλους χρήστες, καθώς και η συμμετοχή τους σε κοινότητες ή δίκτυα για ενίσχυση της διαπολιτισμικής συνείδησης.

- **Content-creation** (Αρχή Δημιουργίας Περιεχομένου):

  Δημιουργία και επεξεργασία νέου εκπαιδευτικού περιεχομένου, είτε μέσω επεξεργαστή κειμένου, είτε επεξεργαστές εικόνων και βίντεο. Διευκόλυνση σε θέματα παραγωγής δημιουργικών εκφράσεων με γνώμονα τη μάθηση, αξιοποιώντας εύκολα και αποτελεσματικά πνευματικά δικαιώματα.

- **Safety** (Αρχή Ασφαλείας):

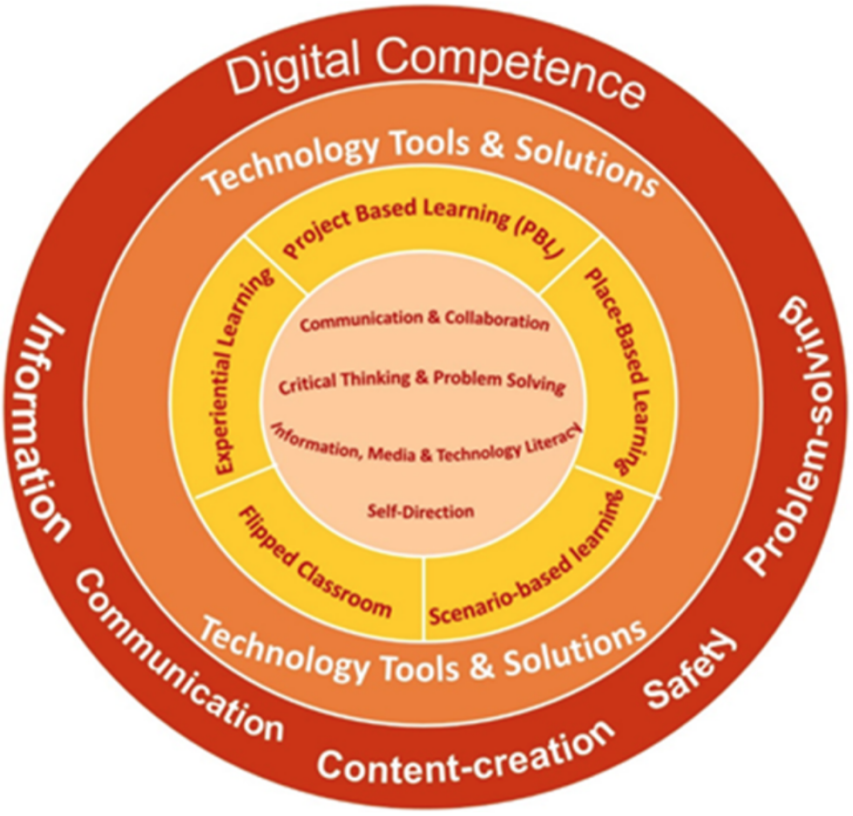
  Έμφαση στην προστασία προσωπικών δεδομένων, ψηφιακής ταυτότητας των χρηστών και διασφάλιση ασφαλούς και βιώσιμης λειτουργίας του συστήματος.

- **Problem-solving** (Αρχή Επίλυσης Προβλημάτων):

  Αναγνώριση και εντοπισμός των ψηφιακών αναγκών των χρηστών με σκοπό την ορθή λήψη αποφάσεων ως προς τα ψηφιακά εργαλεία που τους διατίθενται. Ταυτόχρονα ελέγχεται πώς μπορούν να διδαχθούν πιο αφηρημένες έννοιες και προβλήματα, με δημιουργικό τρόπο, μέσω αυτών των ψηφιακών εργαλείων.

Οι παραπάνω αρχές – επιθυμητές δεξιότητες – ενισχύονται από την ενσωμάτωση εκπαιδευτικών μεθόδων ενώ η συμπληρωματική χρήση των ψηφιακών εργαλείων, αναδεικνύει το επιθυμητό αποτέλεσμα και τον σκοπό του προγράμματος.

Το σχήμα που ακολουθεί δείχνει τον πυρήνα της ιδέας του Up2U προγράμματος και τα επί μέρους τμήματα από τα οποία απαρτίζεται. Ξεκινώντας από τον εσωτερικό κύκλο, οι *Απαραίτητες δεξιότητες ενός φοιτητή στον 21ο αιώνα*, που αποτελούν και τον επιθυμητό στόχο. Αυτές οι δεξιότητες θα πρέπει να αρχίσουν να καλλιεργούνται, ήδη, από τη δευτεροβάθμια εκπαίδευση. Εφαρμόζοντας, λοιπόν, *Εκπαιδευτικές Μεθόδους* όπου ο μαθητής κατέχει κεντρικό ρόλο (learner-centered education), όπως Μάθηση μέσω Έργου (Project Based Learning), Βιωματική Μάθηση (Experiential Learning), Αντεστραμμένη τάξη (Flipped Classroom), Μάθηση βάσει Σεναρίου (Scenario-based Learning), Μάθηση βάσει Τόπου (Place-Based Learning), ο εκπαιδευτικός επιχειρεί να αναδείξει αυτές τις δεξιότητες σε κάθε μαθητή. Ο τρίτος κύκλος, *Τεχνολογικά εργαλεία και λύσεις*, αποτελεί το μέσο που χρησιμοποιεί ο εκπαιδευτικός και «ντύνει», με τρόπο συμπληρωματικό, τις εκάστοτε εκπαιδευτικές μεθόδους. Όλα τα παραπάνω περιλαμβάνονται, τελικά, στις *Ψηφιακές Ικανότητες*, ή αλλιώς τις *Αρχές του Up2U*, που εξασφαλίζουν τη συνοχή όλων των επιπέδων και ανταποκρίνονται στους στόχους που έχουν τεθεί εξ΄ αρχής.

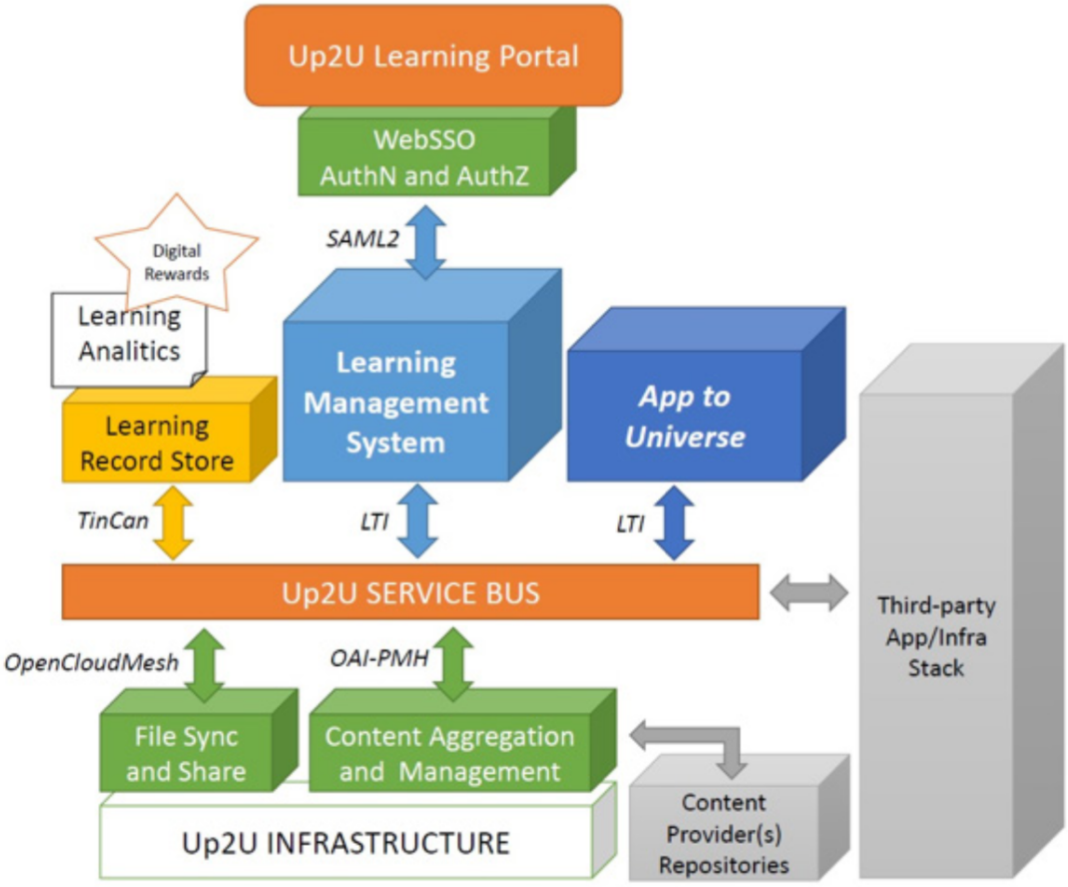


Σχήμα 1.1: Αρχές του Up2U και εκπαιδευτικές μέθοδοι

1.4 Αρχιτεκτονική και Σχεδιασμός του Up2U

Για τον ποικιλόμορφο τρόπο λειτουργίας του, το Up2U, αποτελείται από διαφορετικά λογισμικά και πρωτόκολλα λειτουργίας που θα πρέπει να αλληλεπιδρούν και να επικοινωνούν μεταξύ τους, χωρίς να παρεμβαίνουν στην εμπειρία του χρήστη. Η αρχιτεκτονική του οικοσυστήματος επομένως, βασίζεται στη διαλειτουργικότητα του συστήματος και συγκεκριμένα σε εργαλεία όπως το Learning Tools Interoperability (LTI) και το Experience API (xAPI). Αυτά επιτρέπουν την επαλήθευση στοιχείων και την ενσωμάτωση συμπληρωματικών εκπαιδευτικών εργαλείων στις κύριες υπηρεσίες της πλατφόρμας και στο LMS, ενώ παράλληλα, ανταλλάσσουν δεδομένα μεταξύ των υπηρεσιών και ενός Learning Record Store (LRS) που αποθηκεύει δεδομένα για περαιτέρω LA μελέτη. Η σύνδεση των χρηστών είναι εφικτή μέσω ελέγχου ταυτοποίησης SSO και άφιξη σε σελίδα της LMS πλατφόρμας Moodle.

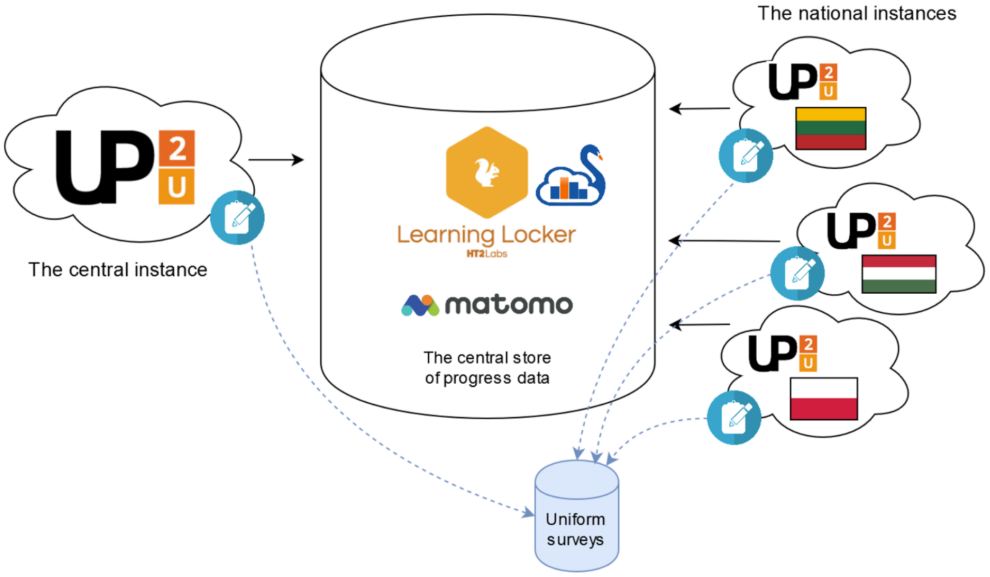
Στο σχήμα που ακολουθεί αναπαρίσταται η αρχιτεκτονική του Up2U Application Toolbox, οι επί μέρους υπηρεσίες και πρωτόκολλα διαλειτουργικότητας τα οποία εφαρμόζονται για την επικοινωνία μεταξύ αυτών.



Σχήμα 1.2: Αρχιτεκτονική του Up2U

1.4.1 Στιγμιότυπα - e-Service

Με την παραπάνω αρχιτεκτονική το Up2U πρόγραμμα εγκατέστησε την πλατφόρμα του σε σχολεία της Ευρώπης που συμμετείχαν σε αυτό. Η αρχική εγκατάσταση του Application Toolbox, ήταν το Κεντρικό Στιγμιότυπο (Central e-Service). Από αυτό, προέκυψαν και τα υπόλοιπα στιγμιότυπα, τα Εθνικά Στιγμιότυπα (National e-Service), προσαρμοσμένα στις ανάγκες και προϋποθέσεις των σχολείων στη Λιθουανία, την Ουγγαρία και την Πολωνία. Τα δεδομένα από τα Στιγμιότυπα της Λιθουανίας, της Πολωνίας και της Ουγγαρίας (Lithuanian, Polish, Hungarian e-Services) συλλέγονται κεντρικά χρησιμοποιώντας το Learning Locker και την εφαρμογή του Matomo ώστε να διασφαλιστεί η συνοχή στο πρόγραμμα και να υπάρχει ομοιογενής αξιολόγηση και βελτίωση.



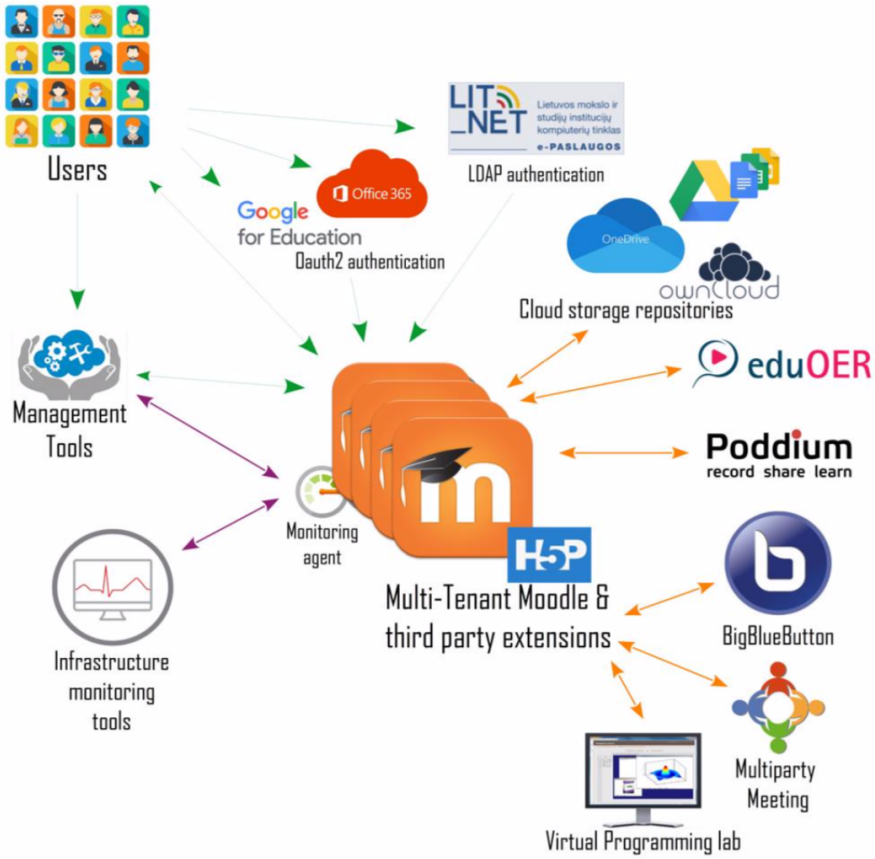
Σχήμα 1.3: Στιγμιότυπα του Up2U

Central e-Service - Κεντρικό Στιγμιότυπο

Το Κεντρικό Στιγμιότυπο αποτελεί τον αρχικό σχεδιασμό του Application Toolbox, πάνω στον οποίο βασίστηκαν και τα υπόλοιπα στιγμιότυπα. Η εφαρμογή του Κεντρικού Στιγμιοτύπου ενεργοποιήθηκε στη Γερμανία, την Ελλάδα, το Ισραήλ, την Ιταλία, την Πορτογαλία και τη Σλοβενία. Όπως αναφέρθηκε προηγουμένως, τα δεδομένα από όλες τις εφαρμογές συλλέγονται κεντρικά. Σε κάθε χώρα όπου το πρόγραμμα τέθηκε σε εφαρμογή προηγήθηκε εκτενής παρουσίαση των λειτουργιών του καθώς και κατάλληλη εκπαίδευση των εκπαιδευτικών που θα το χρησιμοποιούσαν.

Lithuanian e-Service - Λιθουανικό Στιγμιότυπο

Το Λιθουανικό Στιγμιότυπο τέθηκε σε λειτουργία το 2018, ύστερα από παρουσιάσεις και εκπαιδεύσεις σε πάνω από 50 σχολεία ενδιαφερόμενα για τη χρήση του LMS Moodle. Σταδιακά, η εφαρμογή της υποδομής του Up2U εξαπλώθηκε σε περισσότερα σχολεία δημιουργώντας την ανάγκη για τη δημιουργία μιας εθνικής Up2U υποδομής, την National Up2U υποδομή. Η εφαρμογή του National Up2U πραγματοποιήθηκε με την συνεργασία του Πανεπιστημίου Τεχνολογίας του Κάουνας (KTU) και του Λιθουανικού Δικτύου Έρευνας και Εκπαίδευσης LITNET.



Σχήμα 1.4: Υπηρεσίες και δομή του Λιθουανικού Στιγμιοτύπου

Όπως φαίνεται και στο Σχήμα 3.5, η δομή του Λιθουανικού Στιγμιοτύπου βασίζεται σε μία Multi-Tenant αρχιτεκτονική, δηλαδή λογισμικό που μπορεί να μοιραστεί σε περισσότερους διακομιστές, βασισμένη στο Moodle. Σε αυτήν προστέθηκαν περισσότερα add-ons του Up2U Application Toolbox, ενώ η χρήση τους ήταν εφικτή μέσω της Single Sign On (SSO) υπηρεσίας.

Polish e-Service - Πολωνικό Στιγμιότυπο

Το Πολωνικό Στιγμιότυπο ξεκίνησε το 2018 ύστερα από συμμετοχή 15 σχολείων στο πρόγραμμα Συνεχούς Επαγγελματικής Εκπαίδευσης (Continuous Professional Development), με τη σταδιακή προσθήκη περισσότερων σχολείων ανά τα έτη. Το ενδιαφέρον για την πλατφόρμα κορυφώθηκε με την ανακήρυξη της πανδημίας του Covid-19 και την επιβολή των περιοριστικών μέτρων. Περισσότερα από 141 λύκεια και 551 καθηγητές ξεκίνησαν να χρησιμοποιούν την Up2U πλατφόρμα.

openUp2U e-Service

Το openUp2U είναι ένα στιγμιότυπο που έγινε διαθέσιμο στην εκπαιδευτική κοινότητα την 25η Μαρτίου 2020. Αποτελεί ένα στιγμιότυπο, συνεργασίας του Up2U και του Mid Sweden University (MIUN) που δημιουργήθηκε ως μια νέα υπηρεσία του Up2U ώστε να εξυπηρετεί όλους τους Ευρωπαίους χρήστες εν μέσω της COVID-19 πανδημίας και τις αυξημένες ζήτησης σε υπηρεσίες τηλεκπαίδευσης.

Βασικές υπηρεσίες του openUp2U ήταν το Moodle, eduMEET, SWAN, CERNBox, όπως αυτά εξηγήθηκαν στην Παράγραφο 1.2.

Κεφάλαιο 2

Web Analytics

2.1 Web Analytics

Ο σύγχρονος τρόπος ζωής είναι συνυφασμένος με εφαρμογές διαδικτύου (web-based applications), οι οποίες σε συνδυασμό με τα πολυχρηστικά εργαλεία του χώρου της Ανάλυσης, αποτελούν μια εντυπωσιακή πηγή πληροφόρησης της συμπεριφοράς των χρηστών. Κάθε φορά που ένας χρήστης εισέρχεται σε μια ιστοσελίδα (webpage), επιλέγει έναν σύνδεσμο (hyperlink), «κατεβάζει» ένα αρχείο (download), ή απλώς πλοηγείται στο διαδίκτυο, αφήνει, πολλές φορές εν αγνοία του, ένα ίχνος. Βλέποντας τις ευκαιρίες που πηγάζουν από αυτό, ο επιχειρηματικός κόσμος έχει στρέψει το ενδιαφέρον του και επενδύει με αμείωτο ρυθμό στο χώρο των Web Analytics.

Ορισμός 1. *Σύμφωνα με το Web Analytics Association (2008), τα Web Analytics ορίζονται ως η μέτρηση, συλλογή, ανάλυση διαδικτυακών πληροφοριών με σκοπό την κατανόηση και βελτιστοποίηση της χρήσης του διαδικτύου.*

Συνεπώς, τα Web Analytics χρησιμοποιούνται ως μέσο απόκτησης πληροφοριών τόσο για την συμπεριφορά του χρήστη, όσο και για τις ανάγκες αυτού στο διαδίκτυο, ώστε να εντοπισθούν σημεία βελτίωσης της εμπειρίας του και να ελαχιστοποιηθούν τα σφάλματα. Υπάρχουν πολλά λογισμικά Web Analytics με δυνατότητα συλλογής τόσο ποσοτικών όσο και ποιοτικών παρατηρήσεων και μπορεί να είναι είτε Ανοιχτά/ Ελεύθερα (Free Open Source (FOSS)), είτε Ιδιοταγή (Proprietary), είτε να παρέχονται απομακρυσμένα ως Υπηρεσία (Service Hosted Software). Καθένας από αυτούς τους τρεις τύπους λογισμικού έχει πλεονεκτήματα και μειονεκτήματα που καθορίζονται από την επιθυμητή χρήση.

Στον παρακάτω πίνακα συνοψίζονται μερικά από τα πιο δημοφιλή λογισμικά που διατίθενται για συλλογή Web Analytics, η εταιρεία που τα έχει δημιουργήσει, ο τύπος τους (αν είναι, δηλαδή, ανοιχτό, ιδιοταγές λογισμικό ή αν παρέχεται ως υπηρεσία) καθώς και μία ενδεικτική τιμή.

Εταιρεία	Λογισμικό	Tracking Method	Είδος Εργαλείου	Τιμή
Google	Google Analytics	Cookies via JavaScript	Hosted Software as service	Free (Standard) 150,000 Annual (Premium)
Matomo (former Piwik)	Matomo	JavaScript, Web log files	FOSS	Free, From 29$ per month (Cloud)
Adobe Systems	Adobe Analytics	Cookies via JavaScript	Hosted Software as service	Negotiable
Yandex	Yandex Metrica	Cookies via JavaScript	Hosted Software as service	Free
Yahoo!	Yahoo!Web Analytics	Cookies via JavaScript	Hosted Software as service	Free
Mixpanel	Mixpanel	Cookies via JavaScript	Hosted Software as service	Free, Monthly plans
StatCounter	StatCounter	Cookies via JavaScript	Hosted Software as service	Free, Monthly plans
Hotjar	Hotjar	Cookies via JavaScript	Hosted Software as service	Free, Business plans
-	Open Web Analytics	JavaScript or PHP pagetag	FOSS	Free
IBM	Tealeaf CX	Network Traffic Monitor	Proprietary	Business plans
Amplitude	Amplitude	Cookies via JavaScript	Hosted Software as service	Free, Business Plans

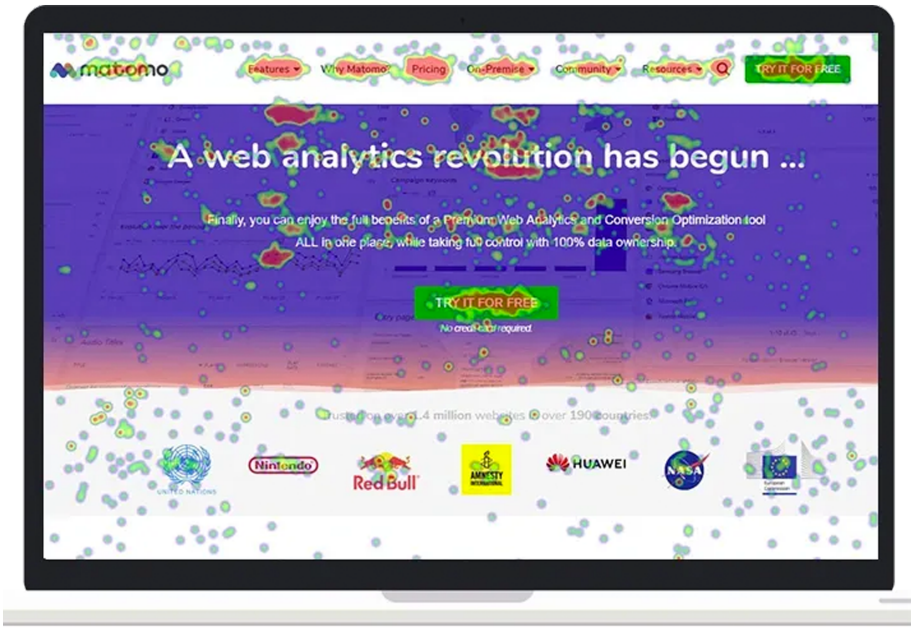
Πίνακας 2.1: Λογισμικά Web Analytics

Από τα παραπάνω, μεγαλύτερο μερίδιο στην αγορά κατέχουν τα Google Analytics ως το δημοφιλέστερο μέσο συλλογής Web Analytics.

2.1.1 Πώς χρησιμοποιούνται τα Web Analytics

1. Βελτιστοποίηση περιβάλλοντος και εμπειρία χρήστη

Γίνεται μέτρηση πληροφοριών σχετικών με την πλοήγηση, το περιεχόμενο και την αλληλεπίδραση του χρήστη με το περιβάλλον της εφαρμογής ή της ιστοσελίδας. Ένα χρήσιμο εργαλείο για αυτό είναι ο χάρτης Heatmap που δείχνει τις περιοχές μεγαλύτερου ενδιαφέροντος και αυτές με τον υψηλότερο ρυθμό σε «κλικ». Ο χάρτης Heatmap δίνει μια γενική εικόνα των περιοχών μεγαλύτερου ενδιαφέροντος και δίνουν μια κατεύθυνση ως προς την ανάλυση που μπορεί να ακολουθηθεί.



Σχήμα 2.1: Παράδειγμα ενός Heatmap από τη σελίδα του Matomo

2. Βελτιστοποίηση ηλεκτρονικού εμπορίου και διατήρηση χρηστών/πελατών

Η κατακόρυφη αύξηση του ηλεκτρονικού εμπορίου, και των εν γένει ηλεκτρονικών υπηρεσιών, έχει καταστήσει ουσιαστική την κατανόηση των τάσεων και των αναγκών των χρηστών-καταναλωτών τους. Πλέον οι χρήστες δέχονται στοχευμένες πληροφορίες, προωθητικές ενέργειες και διαφημίσεις βάσει των ενδιαφερόντων τους και μετράται η αλληλεπίδραση του χρήστη με το περιβάλλον της συναλλαγής. Τελικός στόχος αυτής της ευρείας χρήσης των Web Analytics είναι η διαχείριση της εμπιστοσύνης των χρηστών-καταναλωτών και η εξασφάλιση πως κάθε επίσκεψη στη σελίδα ακολουθείται και από μια ενέργεια όπως είναι η πραγματοποίηση συναλλαγής, η επικοινωνία με κάποιον ειδικό, λήψη αρχείου κ.ά..

3. Καταγραφή των πηγών εισόδου και μέτρηση της επιτυχίας των προωθητικών ενεργειών

Ερωτήσεις όπως «Ποια πηγή εισόδου χρησιμοποίησαν οι χρήστες και από πού βρήκαν πληροφορίες για την ιστοσελίδα;», ή «Ποιες συσκευές χρησιμοποίησαν για την είσοδο στη σελίδα;», καθώς και γενικότερη συσχέτιση των πηγών εισόδου και των ενεργειών ή κόστους μάρκετινγκ προωθήσεων, είναι σημαντικοί παράγοντες μελέτης για την σύγκριση Καίριων Δεικτών Απόδοσης (Key Performance Indicators (KPI) και εύρεση τρόπων βελτίωσής τους.

4. Εντοπισμός προβλημάτων και βελτίωση της απόδοσης των ηλεκτρονικών εφαρμογών

Ο εντοπισμός σφαλμάτων μπορεί πλέον να γίνεται είτε μελετώντας την απόδοση της ηλεκτρονικής εφαρμογής σε πραγματικό χρόνο, είτε αξιοποιώντας ιστορικά δεδομένα. Για την διαδικασία αυτή είναι χρήσιμες οι μετρικές σχετικές με το μέσο όρο φόρτωσης του προγράμματος περιήγησης, η γεωγραφική τοποθεσία του εκάστοτε χρήστη, η απόδοση των τεχνολογικών συστημάτων σε εφαρμογή και άλλες παράμετροι που μπορεί να επηρεάζουν την ομαλή και βέλτιστη εμπειρία των χρηστών.

Αναλύοντας μετρικές που αφορούν τα παραπάνω και κατανοώντας τον τρόπο λειτουργίας μιας ηλεκτρονικής πλατφόρμας, είναι εφικτή η βελτιστοποίηση των παρεχόμενων υπηρεσιών, η προσέλκυση νέων χρηστών ή διατήρησης των υπαρχόντων, και η γενικότερη ανάπτυξη προς όφελος τόσο των επισκεπτών όσο και της υπηρεσίας που παρέχει τις ενέργειές της.

2.1.2 Μετρικές Web Analytics

Η βασικότερη ανάλυση που πραγματοποιείται με τα Web Analytics είναι αυτή που χρησιμοποιεί μετρικές όπως ο αριθμός των επισκέψεων σε έναν ιστότοπο, τη διάρκεια της επίσκεψης. Αυτή η πρώτη ανάλυση αποτελεί την αρχή άλλων αναλύσεων που τη διαδέχονται καθώς και την περαιτέρω ανάπτυξη πορισμάτων.

Λόγου χάρη, η συλλογή και ανάλυση μιας συγκεκριμένης μετρικής κατά τη διάρκεια μιας δεδομένης χρονικής περιόδου, μπορεί να αξιοποιηθεί για μια μεταγενέστερη Ανάλυση Τάσεων, ή για μια Ανάλυση Διασποράς με σκοπό την κατανόηση της ποικιλομορφίας των χρηστών και, συνεπώς, την κατανομή τους σε συστάδες. Παρομοίως, βασιζόμενοι στις πρωταρχικές μετρικές, μπορεί να διεξαχθεί μια Εφαρμοσμένη Ανάλυση Συμπεριφοράς για την κατανόηση των αναγκών και επιλογών των επισκεπτών σε μία ηλεκτρονική σελίδα.

Βασικές Μετρικές που χρησιμοποιούνται στα Web Analytics:

Μετρικές Web Analytics	
Περιγραφής επισκέψεων	- Αριθμός επισκέψεων (Visits) - Διάρκεια επίσκεψης (Visit Duration) - Σελίδες εισόδου/ εξόδου - Όροι αναζήτησης - Click-through-rate - Πηγή επισκεψιμότητας (Source of Entry) - Σύνδεσμοι που επιλέχθηκαν - Χρόνος φόρτωσης σελίδας
Περιγραφής χρηστών	- Νέος επισκέπτης (New Visit) - Επανερχόμενος επισκέπτης (Returning Visit) - Αριθμός επισκέψεων ανά επισκέπτη - Συχνότητα επισκέψεων - Συχνότητα μιας ενέργειας - Γεωγραφικές τοποθεσίες χρηστών
Συμπεριφοράς χρηστών	- Ποσοστό εγκατάλειψης (Bounce Rate(%)) - Λήψεις αρχείου (Downloads) - Σύνολο σελίδων που επισκέφθηκαν (Pageviews) - Σελίδες επίσκεψης ανά χρήστη - Αριθμός χρηστών που μεταφόρτωσε αρχείο - Ρυθμός κύλισης δρομέα (Scrolling Progression)
Μετατροπών	- Αριθμός επιτυχίας στόχων (conversion)

Πίνακας 2.2: Μετρικές Web Analytics

Μετρική Bounce Rate(%)

Μία μετρική που χρησιμοποιείται ευρέως σε όλους τους κλάδους που αποθηκεύουν και αναλύουν Web Analytics αποτελεί η Bounce Rate(%). Η μετρική αποτυπώνει το ποσοστό των επισκεπτών που εισέρχονται σε μια ιστοσελίδα χωρίς να αλληλεπιδράσουν περαιτέρω με αυτήν και το υλικό της και εξέρχονται χωρίς περαιτέρω ενέργειες, δηλαδή όταν ένας χρήστης «κάνει bounce». Αυτό μπορεί να συμβεί είτε πατώντας σύνδεσμο για κάποια άλλη ιστοσελίδα, είτε κλείνοντας το παράθυρο πλοήγησης, είτε πληκτρολογώντας μία νέα διεύθυνση, επιστρέφοντας στην σελίδα αρχικής αναζήτησης ή εάν τερματιστεί ο χρόνος παραμονής.

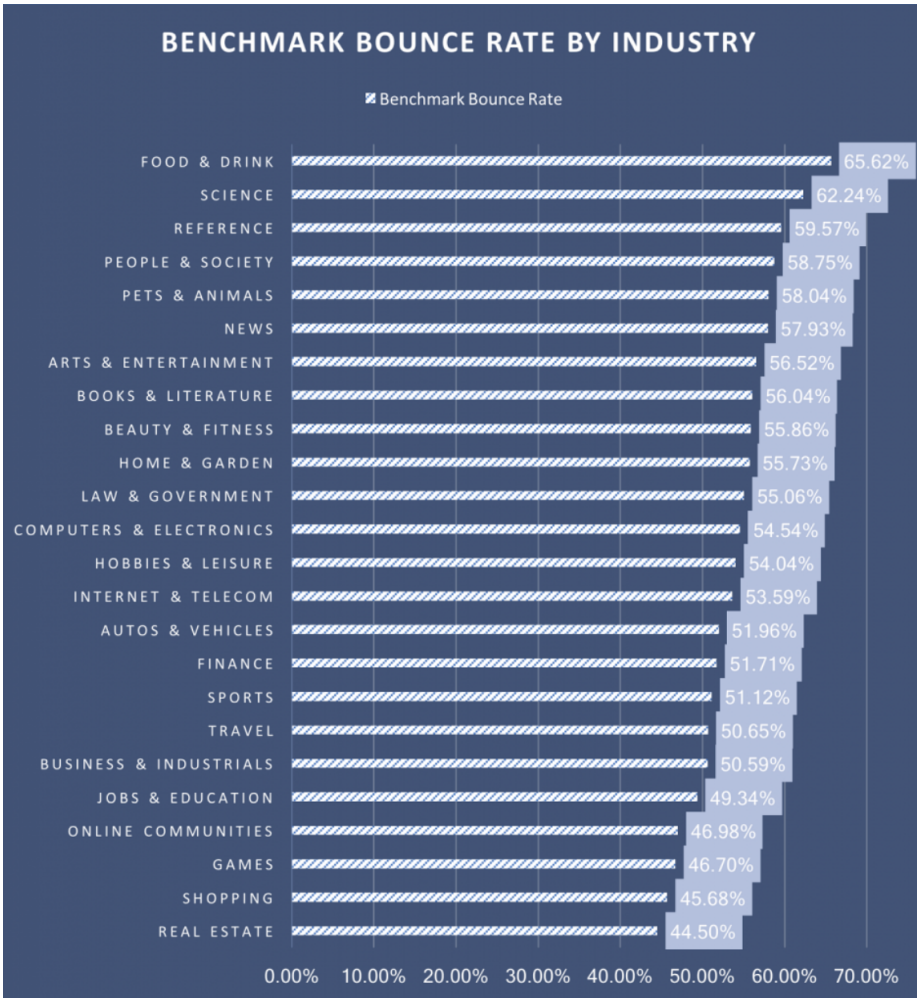
Υπολογίζεται ως εξής:

$$BounceRate(\%) = \frac{\begin{array}{c}\text{Σύνολο επισκεπτών σε μοναδική σελίδα του ιστοτόπου} \\ \text{που εξέρχονται χωρίς να κάνουν περαιτέρω ενέργεια}\end{array}}{\text{Σύνολο επισκεπτών στη σελίδα}} \times 100 \quad (2.1)$$

Ένα υψηλό Bounce Rate(%) μπορεί να θεωρηθεί αρνητικό, αντίστοιχα ένα χαμηλό Bounce Rate(%) θετικό, για την προσέλκυση επισκεπτών στον επιθυμητό διαδικτυακό χώρο ή πλατφόρμα, ωστόσο αυτό είναι μια απλούστευση της πραγματικότητας. Για να μπορέσει κανείς να μελετήσει τη μετρική και να καταλήξει στο εάν ένα υψηλό Bounce Rate(%) είναι αρνητικό θα πρέπει να γνωρίζει το πλαίσιο στο οποίο ανήκει, την αγορά στην οποία ανταποκρίνεται και το περιεχόμενο της σελίδας. Ένα υψηλό Bounce Rate(%) μπορεί να δείχνει πως το περιεχόμενο δεν κατάφερε να κρατήσει το ενδιαφέρον του κοινού ή υπάρχει κάτι που χρήζει βελτίωσης, μπορεί ωστόσο να δείχνει πως οι επισκέπτες βρήκαν επαρκείς τις πληροφορίες και το υλικό που διατίθεται ή πως απλώς η υπηρεσία προσφέρει αυτή τη μοναδική σελίδα.

Η εικόνα του Σχήματος 2.2 δείχνει έναν μέσο όρο Bounce Rate(%) για διαφορετικούς κλάδους.

Φαίνεται πως στον χώρο της Εστίασης (Food & Drink) ο μέσος όρος του Bounce Rate(%) είναι 65.62%, που σημαίνει πως είτε οι σελίδες στον χώρο αυτόν περιέχουν όλη την πληροφορία που αναζητά ο επισκέπτης (μενού, τηλέφωνο επικοινωνίας, διεύθυνση), είτε δεν προσφέρουν κάποια άλλη ενέργεια, εξού και το πιο υψηλό Bounce Rate(%). Αντίθετα για τον κλάδο Εργασία και Εκπαίδευση (Jobs & Education) το Bounce Rate(%) είναι 49.34%, που ενδεχομένως να αντικατοπτρίζει το ότι σχεδόν οι μισοί πλοηγούνται και σε άλλες σελίδες των διαθέσιμων ιστοσελίδων, πλατφορμών στις οποίες εισέρχονται για να αναζητήσουν κι άλλο υλικό.



Σχήμα 2.2: Μέσος όρος Bounce Rate(%) διαφορετικών κλάδων

Πηγή: https://cxl.com/guides/bounce-rate/benchmarks/

Γνωρίζοντας από πού προέρχονται τα δεδομένα, το κανάλι εισόδου των χρηστών (Direct Entry, Search Engines, Social Networks, Websites), καθώς και τη συσκευή από την οποία εισήλθαν οι χρήστες, μπορεί να υπάρξει ολοκληρωμένη εικόνα για το Bounce Rate(%) κι επομένως για την στρατηγική που μπορεί να ακολουθήσει κανείς, τόσο για τον σχεδιασμό του εικονικού περιβάλλοντος, όσο και για τον τρόπο με τον οποίο θα κάνει το περιεχόμενό του προσιτό και ενδιαφέρον για τους επισκέπτες.

Στη συνέχεια αυτής της διπλωματικής, θα γίνει μελέτη ορισμένων από τις παραπάνω μετρικές που αφορούν τον τύπο επισκεπτών (Visitor Type Metrics), τις ενέργειές τους (Activity Metrics), ενώ σημαντικό μέρος της μελέτης αφορά την ταξινόμηση του Bounce Rate(%) σε χαμηλό/ υψηλό, με στόχο να κατανοηθεί ο τρόπος χρήσης της πλατφόρμας του Up2U όπως εγκαταστάθηκε στην Πολωνία.

Κεφάλαιο 3

Θεωρητικό Υπόβραθρο

3.1 Καθαρισμός & Προεπεξεργασία Δεδομένων

Ένα σύνηθες πρόβλημα μετά την εξόρυξη δεδομένων είναι η ανάγκη καθαρισμού των Data Frame από σφάλματα, μη επιθυμητούς τύπους δεδομένων ή ελλείψεις τιμών. Αυτά μπορεί να προκύπτουν λόγω ανθρώπινου, μηχανικού σφάλματος ή τρόπου συλλογής των δεδομένων.

Επομένως, πριν την ανάλυση και επεξεργασία δεδομένων προηγείται η προετοιμασία και ο καθαρισμός τους.

3.1.1 Χειρισμός Ελλιπών Τιμών

Η απουσία τιμών είναι από τα βασικά σφάλματα που θα πρέπει να αντιμετωπιστούν. Σε σχέση με τα υπόλοιπα σφάλματα αποτελεί σημαντικό εμπόδιο στην επεξεργασία δεδομένων καθώς, τις περισσότερες φορές, είναι αδύνατο να ανακτηθούν οι τιμές αυτές κι, ως εκ τούτου, ο αναλυτής να πρέπει να επέμβει με τρόπο που θα επηρεάσει, αναγκαστικά, τα αποτελέσματα σε μικρό βαθμό. Σε αυτή την περίπτωση σκοπός είναι η όσο το δυνατό μικρότερη εισαγωγή μεροληψίας στα δεδομένα, με τη βέλτιστη μέθοδο αντιμετώπισης να εξαρτάται από τον τύπο-μηχανισμό των ελλιπών τιμών.

Υπάρχουν τρία είδη μηχανισμών ελλιπών τιμών (Donald B.Rubin, 1987):

- Εντελώς Τυχαία Έλλειψη - Missing Completely At Random (MCAR)

  Έστω μία μεταβλητή Y που περιέχει ελλιπή δεδομένα, και μια άλλη μεταβλητή X που παρατηρείται πάντα. Λέμε πως τα δεδομένα ακολουθούν Εντελώς Τυχαία Έλλειψη Missing Completely At Random όταν η πιθανότητα μιας παρατήρησης να λείπει δεν εξαρτάται ούτε από τις ίδιες τις μεταβλητές Y που περιέχουν ελλιπείς τιμές αλλά ούτε και από τις μεταβλητές X των οποίων τα δεδομένα παρατηρούνται κανονικά.

$$P(R = 1|X, Y) = P(R = 1) \tag{3.1}$$

όπου

$$R = \begin{cases} 1, Ymissing \\ 0, Yobserved \end{cases}$$

MCAR θεωρείται, για παράδειγμα, μία ελλιπής παρατήρηση από ένα σύνολο εργαστηριακών ελέγχων λόγω τυχαίας απώλειας του δείγματος. Στην πράξη, οι Missing Completely At Random τιμές αποτελούν μία σπάνια περίπτωση καθώς συνήθως υπάρχουν συσχετίσεις μεταξύ των τιμών και των μεταβλητών.

- Τυχαία Έλλειψη - Missing At Random (MAR)

  Στην Τυχαία Έλλειψη τα δεδομένα δεν λείπουν εντελώς τυχαία, όπως την προηγούμενη κατηγορία, αλλά η πιθανότητα έλλειψης κάποιας παρατήρησης μπορεί να εξαρτάται από τις παρατηρούμενες τιμές όχι όμως από την τιμή αυτή καθεαυτή. Έστω, δηλαδή, μία μεταβλητή Y που περιέχει ελλιπείς τιμές, και έστω μία μεταβλητή X στην οποία δεν παρατηρούνται ελλιπείς τιμές. Λέμε πως οι ελλιπείς τιμές στην Y είναι Missing At Random εάν η πιθανότητα της Y να λείπει δεν εξαρτάται από τον εαυτό της, δεδομένου πως η X έχει οριστεί και μπορεί να ελεγχθεί.

  $$P(R = 1|X, Y) = P(R = 1|X) \tag{3.2}$$

  Δηλαδή, έλλειψη στην Y μεταβλητή εξαρτάται από άλλες παρατηρούμενες μεταβλητές. Παραδείγματος χάριν, σε ιατρική έρευνα για την αρτηριακή πίεση και επίπεδα χολιστερόλης, εάν σε άτομα με λιγότερους ιατρικούς ελέγχους λείπουν συχνότερα οι τιμές τις πίεσης, τότε ο μηχανισμός θεωρείται MAR. Η πιθανότητα μη καταγραφής οφείλεται στην μεταβλητή «ιατρικός έλεγχος» κι όχι στην τιμή της αρτηριακής πίεσης.

- Μη Τυχαία Έλλειψη - Missing Not At Random (MNAR)

  Οι τιμές είναι NMAR, όταν η πιθανότητα έλλειψης τιμής μιας μεταβλητής εξαρτάται από την τιμή αυτή καθεαυτή, π.χ. όταν ένας σένσορας καταγραφής θερμότητας έχει κατώφλι που δεν καταγράφει πλέον τιμές εάν ξεπεραστεί.

Σε ένα σύνολο δεδομένων είναι πιθανό να υπάρχουν περισσότεροι από έναν μηχανισμό ελλιπών τιμών, γι΄ αυτό απαιτείται καλή κατανόηση των μεταβλητών και των σχέσεων μεταξύ τους. Για την αντιμετώπιση ελλιπών τιμών μπορούν να εφαρμοστούν διάφορες μέθοδοι ανάλογα τα δεδομένα, την σχέση μεταξύ τους και του είδους του μηχανισμού των ελλιπών τιμών. Ενδεικτικά, μερικές μέθοδοι είναι: απαλοιφή στηλών ή γραμμών, αντικατάσταση με μεθόδους μέσου όρου, αλγόριθμοι μηχανικής μάθησης κ.α..

3.1.2 k-Nearest Neighbors Imputation

Ο αλγόριθμός k-Πλησιέστερων Γειτόνων, k-Nearest Neighbors (kNN), είναι μία μη-παραμετρική, Επιβλεπόμενη Μέθοδος Μηχανικής Μάθησης (Supervised Learning) και χρησιμοποιείται τόσο για κατηγοριοποίηση παρατηρήσεων (Classification) όσο και για προβλήματα Παλινδρόμησης (Regression). Ο αλγόριθμος αξιοποιεί την ιδέα πως όμοιες παρατηρήσεις έχουν όμοια αποτελέσματα, επομένως, μπορούν να κατηγοριοποιηθούν μαζί ή μπορούν να οδηγήσουν σε μία πρόβλεψη βάσει των k πλησιέστερων και όμοιων γειτόνων.

Η kNN Imputation μέθοδος εφαρμόζει τον αλγόριθμο σε ένα σύνολο δεδομένων εντοπίζοντας τις παρατηρήσεις που παρουσιάζουν ομοιότητα, και, εν συνεχεία, εκτιμώντας και αντικαθιστώντας τις τιμές που λείπουν στο σύνολο αυτό. Στην περίπτωση συνεχών τιμών, ο πιο συνηθισμένος τρόπος εντοπισμού των k-γειτόνων είναι μέσω του υπολογισμού της Ευκλείδειας Απόστασης:

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Σημαντικός παράγοντας για την επιτυχία της μεθόδου είναι η επιλογή του αριθμού των k-γειτόνων.

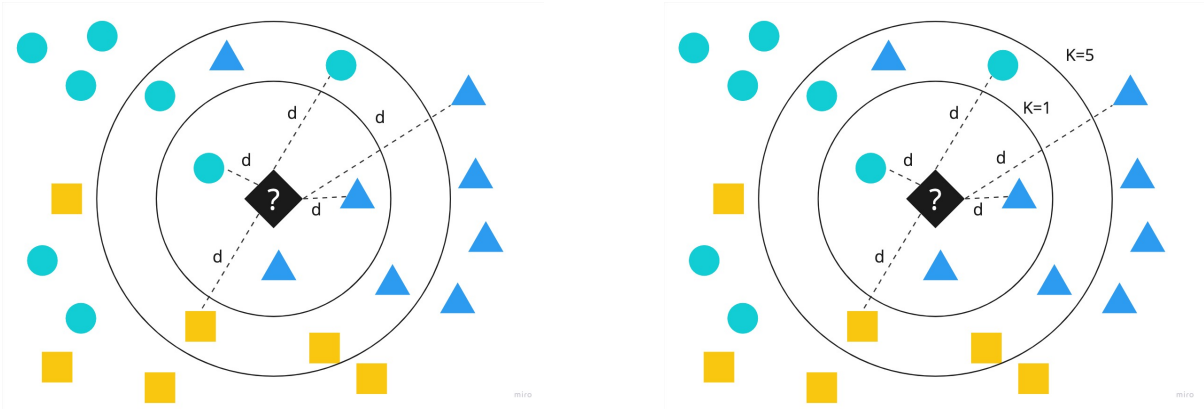
- μικρός αριθμός k → μεγάλος θόρυβος στις τιμές & γρήγορη εκτέλεση
- μεγάλος αριθμός k → ακριβέστερες τιμές & επιβραδυμένη εκτέλεση

Η επιλογή γίνεται βάσει αξιολόγησης των διαφορετικών τιμών του k που δίνουν την ακριβέστερη εκτίμηση. Κρίνεται, λοιπόν, απαραίτητη η πλήρης κατανόηση των δεδομένων και του πεδίου της μελέτης. Παρακάτω αναγράφεται ο αλγόριθμος του kNN Imputation.

Αλγόριθμος: k-Nearest Neighbors

1: Εντοπισμός ελλιπών τιμών στο σύνολο δεδομένων
2: Επιλογή τιμής k-γειτόνων που θα ληφθούν υπόψη για την εκτίμηση
3: Υπολογισμός αποστάσεων μεταξύ άγνωστης τιμής και όλων των παρατηρήσεων με τη χρήση μετρικής απόστασης, συνήθως της Ευκλείδειας Απόστασης
4: Εντοπισμός k-πλησιέστερων γειτόνων
5: Εκτίμηση τιμών βάσει μέσου όρου των k-πλησιέστερων γειτόνων

Τα βήματα επαναλαμβάνονται ξεχωριστά για κάθε ελλιπή τιμή στο σύνολο δεδομένων.



(α΄) Υπολογισμός αποστάσεων (β΄) Εντοπισμός k-πλησιέστερων γειτόνων

Σχήμα 3.1: kNN Imputation

Πλεονεκτήματα και Μειονεκτήματα χρήσης kNN Imputation

Πλεονεκτήματα	**Περιορισμοί**
Απλή και αποτελεσματική μέθοδος	Αργή μέθοδος για μεγάλες βάσεις δεδομένων
Εφαρμογή για ποσοτικές και ποιοτικές μεταβλητές	Η επιλογή του k επηρεάζει σημαντικά τα αποτελέσματα
Δεν απαιτεί δημιουργία κάποιου μοντέλου πρόβλεψης	

Για διευκόλυνση εφαρμογής της, υπάρχουν ανοικτού τύπου λογισμικά στις βιβλιοθήκες μηχανικής μάθησης γλωσσών προγραμματισμού. Για τη γλώσσα Python υπάρχει η βιβλιοθήκη scikit-learn και το αντίστοιχο αντικείμενό της sklearn.impute.KNNImputer. Οι βιβλιοθήκες αυτές και τα αντικείμενα θα χρησιμοποιηθούν κατά τη διάρκεια επεξεργασίας των δεδομένων της παρούσας διπλωματικής.

3.2 Μοντέλα Μηχανικής Μάθησης

Με την ραγδαία αύξηση πληροφορίας, δεδομένων και των σχετικών δυσκολιών που προκύπτουν ως προς την αποθήκευση, συλλογή, οργάνωση και διαχείριση των δεδομένων ο κλάδος της Στατιστικής έχει να αντιμετωπίσει νέες προκλήσεις. Οι προκλήσεις αυτές, η ανάγκη για μάθηση από τα δεδομένα (learning from data), και η διευκόλυνση που προσφέρουν οι υπολογιστές έχουν στρέψει τις στατιστικές επιστήμες στην Επιστήμη των Υπολογιστών και στη Μηχανική Μάθηση (Machine Learning). Οι δύο επιστήμες πλέον αλληλοσυμπληρώνονται προσφέροντας μεγαλύτερη άνεση και ακρίβεια στους αναλυτές.

Για τη Μηχανική Μάθηση συναντώνται συνήθως οι δύο παρακάτω ορισμοί:

Ορισμός 2. *Μηχανική Μάθηση είναι το ερευνητικό πεδίο που δίνει την δυνατότητα στους υπολογιστές να μαθαίνουν χωρίς να είναι αναλυτικά προγραμματισμένοι. - Arthur Samuel, 1959*

Ορισμός 3. *Ένα πρόγραμμα μαθαίνει από μια εμπειρία E ως προς κάποια ενέργεια T και μια απόδοση P, εάν η απόδοση στο T, όπως μετράται από την P, βελτιώνεται με χρήση της εμπειρίας E. - Tom Mitchell, 1977*

Υπάρχουν τριών ειδών προβλήματα στη μηχανική μάθηση:

- **Επιβλεπόμενη Μάθηση - Supervised Learning**

  Τα πιθανά αποτελέσματα είναι γνωστά και αφορούν κάποια πρόβλεψη βάσει των σχέσεων και των μοτίβων των δεδομένων. Τα προβλήματα τύπου Supervised Learning χωρίζονται σε δύο κατηγορίες, τα προβλήματα Παλινδρόμησης (Regression) και τα προβλήματα Ταξινόμησης (Classification).

- **Μη Επιβλεπόμενη Μάθηση - Unsupervised Learning**

  Στα προβλήματα Unsupervised Learning τα πιθανά αποτελέσματα δεν είναι γνωστά. Στόχος είναι περιγραφή των σχέσεων και μοτίβων στα δεδομένα εισόδου που προκύπτουν από συστάδες.

- **Ενισχυτική Μάθηση - Reinforced Learning**

  Τα προβλήματα αφορούν την εκμάθηση μέσω της αλληλεπίδρασης των ενεργειών με το περιβάλλον με στόχο την βελτιστοποίηση των αποφάσεων.

3.2.1 Μοντέλα Ταξινόμησης

3.2.1.1 Logistic Regression

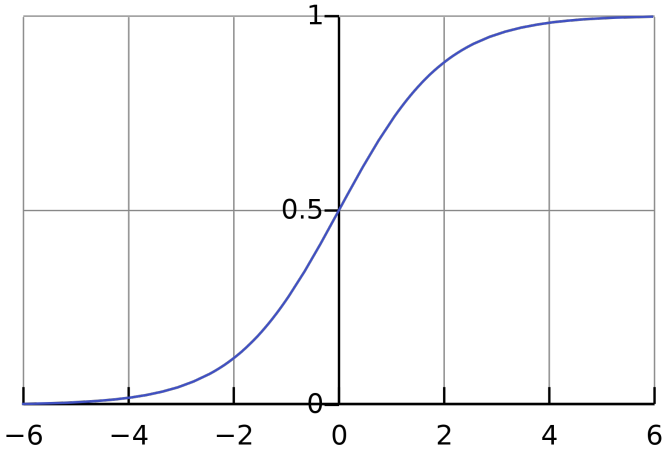
Η λογιστική παλινδρόμηση (Logistic Regression) αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων και ένα μοντέλο επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιείται για προβλήματα ταξινόμησης μιας επεξηγηματικής μεταβλητής y. Βασιζόμενο στη θεωρία πιθανοτήτων, το μοντέλο εκτιμά την πιθανότητα εμφάνισης ενός γεγονότος, π.χ. πιθανότητα εμφάνισης ή μη μιας εκφυλιστικής νόσου. Εάν η πιθανότητα της εκτιμημένης τιμής είναι πάνω από 50%, τότε το μοντέλο προβλέπει σε ποια κλάση ανήκει το στιγμιότυπο. Αυτή η περίπτωση αποτελεί την δυαδική λογιστική παλινδρόμηση binary.

Η συνάρτηση της δυαδικής λογιστικής παλινδρόμησης έχει τη μορφή:

$$p(t) = \frac{e^t}{1+e^t} = \frac{1}{1+e^{-t}} \tag{3.3}$$

$$\mathrm{t} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Η συνάρτηση ακολουθεί σιγμοειδή κατανομή (Sigmoid) και το εύρος του συνόλου τιμών είναι είναι μεταξύ 0 και 1.



Σχήμα 3.2: Σιγμοειδής Καμπύλη

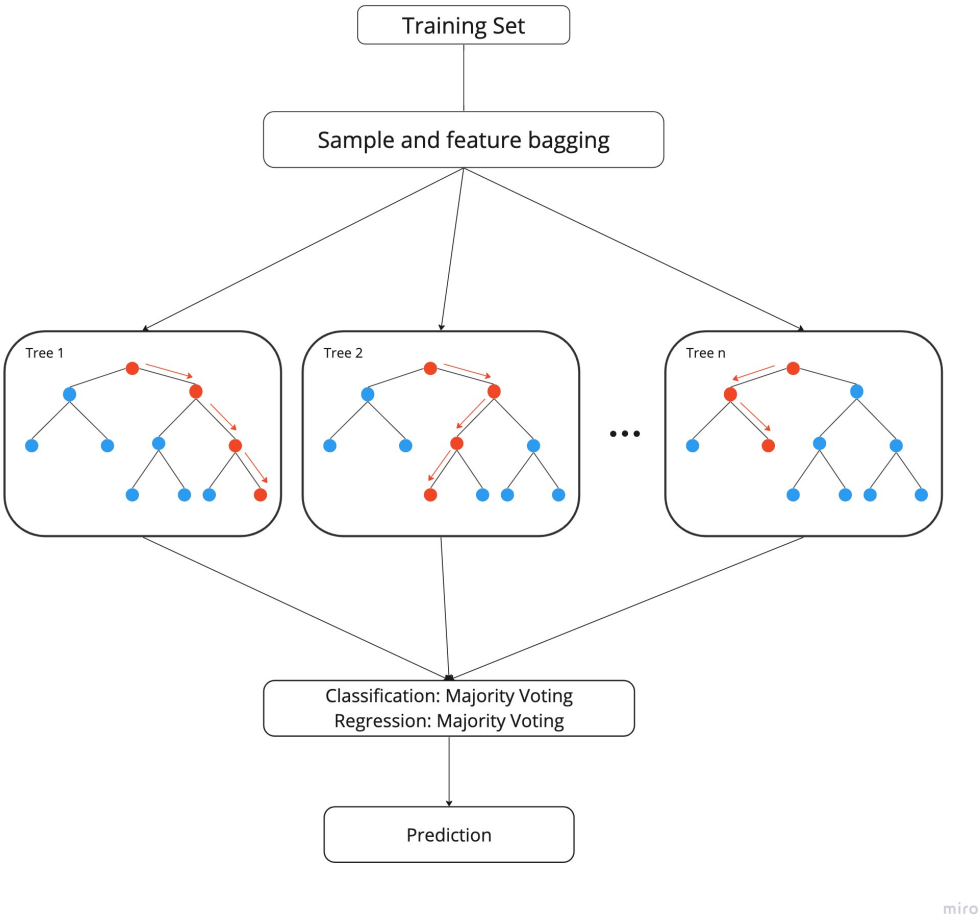
Όταν το μοντέλο έχει εκτιμήσει τους συντελεστές των $x_1, x_2, ..., x_k$ και το εκφράζει ως $\hat{p} = h_\theta(x) = \sigma(x^T\theta)$, δηλαδή τη πιθανότητα μία επεξηγηματική μεταβλητή να ανήκει στην θετική κλάση, τότε μπορεί να κάνει την ταξινόμηση.

$$\hat{y} = \begin{cases} 0 & \text{για} \quad \hat{p} < 0.5 \\ 1 & \text{για} \quad \hat{p} \geq 0.5 \end{cases}$$

Η λογιστική παλινδρόμηση έχει το πλεονέκτημα πως για δυαδικά/ διωνυμικά δεδομένα δίνει τη δυνατότητα ερμηνείας των επεξηγηματικών μεταβλητών, βοηθώντας στην κατανόηση εκείνων που συνεισφέρουν περισσότερο στο μοντέλο.

3.2.1.2 Random Forest

Τα Τυχαία Δάση ή Random Forest(RF), είναι ένα επιβλεπόμενο μοντέλο μηχανικής μάθησης τόσο για προβλήματα Classification όσο και για Regression. Ο αλγόριθμος δημιουργεί υποσύνολα τυχαίων και ανεξάρτητων δέντρων αποφάσεων και βασιζόμενος στη μέθοδο παράλληλου συνδυασμού bagging προσπαθεί να δώσει την υψηλότερη δυνατή απόδοση και σταθερότητα. Το επόμενο σχήμα δείχνει τη λειτουργία του αλγορίθμου.



Σχήμα 3.3: Random Forest

Αλγόριθμος: Random Forest

Έστω $D = (x_1, y_1), \ldots, (x_n, y_n)$ σύνολο εκπαίδευσης, με $x_i = (x_{i,1}, \ldots, x_{i,p})^T$. Για $j = 1, 2, \ldots, k$:

1: Με μέθοδο bootstrap επίλεξε δείγμα D_j μεγέθους n από το D
2: Δημιούργησε τυχαίο δέντρο T_b στο δείγμα, επαναλαμβάνοντας διαδοχικά τα ακόλουθα βήματα για κάθε τελικό κόμβο του δέντρου μέχρι να επιτευχθεί ο ελάχιστος αριθμός κόμβων n_{min}.
   (i) Επίλεξε τυχαία m μεταβλητές από το σύνολο των p μεταβλητών
   (ii) Διάλεξε την καλύτερη μεταβλητή/ σημείο διαχωρισμού από τις m
   (iii) Διάσπασε τον κόμβο σε δύο θυγατρικούς κόμβους
3: Επίστρεψε το σύνολο των δέντρων που δημιουργήθηκε $(T_b)_1^B$

Για πρόβλεψη ενός νέου σημείου x:

Regression:$\hat{f}_{rf}^B = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Classification: Έστω $\hat{C}_b(x)$ η προβλεπόμενη κλάση του b-οστού Random Forest. Τότε $\hat{C}_{rf}^B = \textit{ψήφος πλειοψηφίας}\left\{\hat{C}_b(x)\right\}_1^B$

Στα πλεονεκτήματα του μοντέλου είναι πως:

- μπορεί να αντεπεξέλθει σε πολυδιάστατα προβλήματα
- διαχειρίζεται και υπολογίζει ελλιπείς τιμές
- απαιτεί μικρό χρόνο εκπαίδευσης και πρόβλεψης
- διαχειρίζεται ποικιλοτρόπως τις διαφορετικές κλάσεις
- λαμβάνει αυτόματα υπόψη ένα γενικό σφάλμα μεροληψίας

Ένα ακόμη βασικό χαρακτηριστικό του Random Forest είναι πως υπολογίζει το πόσο σημαντική είναι κάθε μεταβλητή, δηλαδή πόσο συνεισφέρει στο μοντέλο. Μετά την εκπαίδευση του μοντέλου κάθε μεταβλητή αξιολογείται και ταξινομείται έτσι ώστε το άθροισμά τους να είναι ίσο με 1.

3.2.1.3 Gradient Boosting και XGBoost

Gradient Boosting

Το Gradient Boosting αποτελεί μία ακόμη μέθοδο συνδυασμού πολλαπλών δέντρων αποφάσεων ensemble που χρησιμοποιείται για προβλήματα παλινδρόμησης και ταξινόμησης. Γενικός τρόπος λειτουργίας ενός Gradient Boosting είναι πως δημιουργεί απλά δέντρα αποφάσεων (weak learners) με σειριακό τρόπο, έτσι ώστε κάθε νέο δέντρο-μοντέλο να εκπαιδεύεται φέροντας διορθώσεις σε σχέση με το προηγούμενο. Κάθε δέντρο μπορεί να κάνει εκτιμήσεις για ένα μέρος μόνο των δεδομένων, επομένως περισσότερα δέντρα προστίθενται ώστε να βελτιωθεί η απόδοσή του.

Αλγόριθμος: Gradient Boosting

Έστω $D = (x_1, y_1), \ldots, (x_n, y_n)$ σύνολο εκπαίδευσης, και $L(y, F(x))$ η συνάρτηση διαφοράς πραγματικής και εκτιμώμενης τιμής

1: Όρισε αρχικό μοντέλο $F_0(x) = \arg\min_\gamma L(y, F(x)) = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$
2: Για $m = 1, ..., M$ δέντρα:

(i) Για $i = 1, 2, ..., N$ υπολόγισε τα ψευδο-υπόλοιπα:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}}$$

(ii) Προσάρμοσε δέντρο αποφάσεων βάσης $h_m(x)$ στα r_{im} εκπαιδεύοντας το σύνολο εκπαίδευσης $\{(x_i, r_{im})\}_{i=1}^{N}$

(iii) Υπολόγισε:

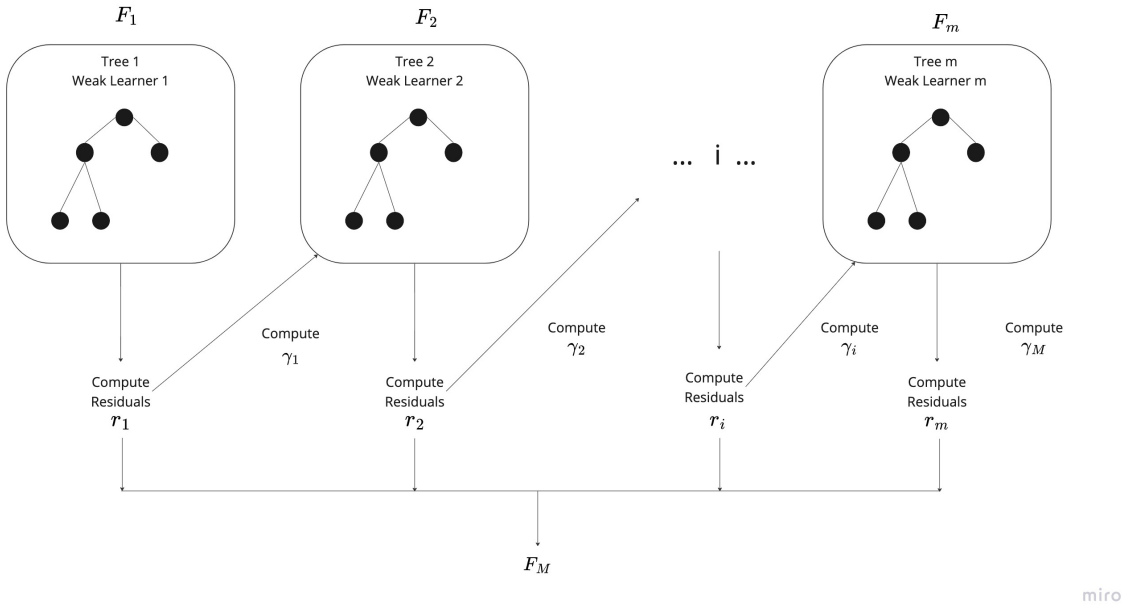
$$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

(iv) Ανανέωσε το μοντέλο

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

3: Τελική πρόβλεψη: $F_M(x)$

Το μοντέλο είναι μη-παραμετρική μέθοδος και μπορεί να αντιμετωπίζει επιτυχώς ακραίες τιμές, έχει υψηλά ποσοστά ακρίβειας, διαθέτει ενσωματωμένη αξιολόγηση των μεταβλητών, και διαχειρίζεται μεγάλο πλήθος πολύπλοκων συνόλων δεδομένων.



Σχήμα 3.4: Απεικόνιση λειτουργίας Gradient Boosting

XGBoost

Το XGBoost ή eXtreme Gradient Boosting αποτελεί μία βελτιστοποιημένη εκδοχή του Gradient Boosting που επεκτάθηκε λόγω του μικρού υπολογιστικού κόστους και της ακρίβειάς του. Υπερτερεί της Gradient Boosting καθώς ενσωματώνει αυτόματα κανονικοποίηση των δεδομένων, διαχειρίζεται ελλιπείς τιμές και πολύπλοκα σύνολα.

3.2.1.4 Support Vector Machine

Μία Μηχανή Διανυσμάτων Υποστήριξης, Support Vector Machine (SVM), είναι ένα ισχυρό μοντέλο μηχανικής επιβλεπόμενης μάθησης που μπορεί να χρησιμοποιηθεί για γραμμικά ή μη γραμμικά προβλήματα ταξινόμησης, παλινδρόμησης και εύρεσης ακραίων τιμών. Τα SVM μοντέλα αναπτύχθηκαν από τον Vladimir Padnik και την ομάδα του το 1995 και έχουν ευρέως χρησιμοποιηθεί για κατηγοριοποίηση γονιδιωμάτων, κειμένων, εικόνων κ.α.. Προσαρμόζονται καλύτερα σε μικρού ή μεσαίου μεγέθους πολύπλοκων δεδομένων και, για τα γραμμικά προβλήματα ταξινόμησης, βασίζονται σε μία συνάρτηση αποφάσεων που λειτουργεί ως όριο υπερεπιπέδου (hyperplane) και χωρίζει τα δεδομένα σε κλάσεις.

Παρακάτω ο τρόπος λειτουργίας του μοντέλου αναλύεται.

Έστω το σύνολο εκπαίδευσης αποτελείται από N ζεύγη $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ με $x_i \in \mathbb{R}^p$ και $y_i \in -1, 1$. Το υπερεπίπεδο ορίζεται ως:

$$w^T x + \beta = 0 \tag{3.4}$$

όπου w διάνυσμα βαρών κάθετο στο επίπεδο και β η απόσταση του υπερεπιπέδου από την αρχή των αξόνων παράλληλα του κανονικού διάνυσμα w.

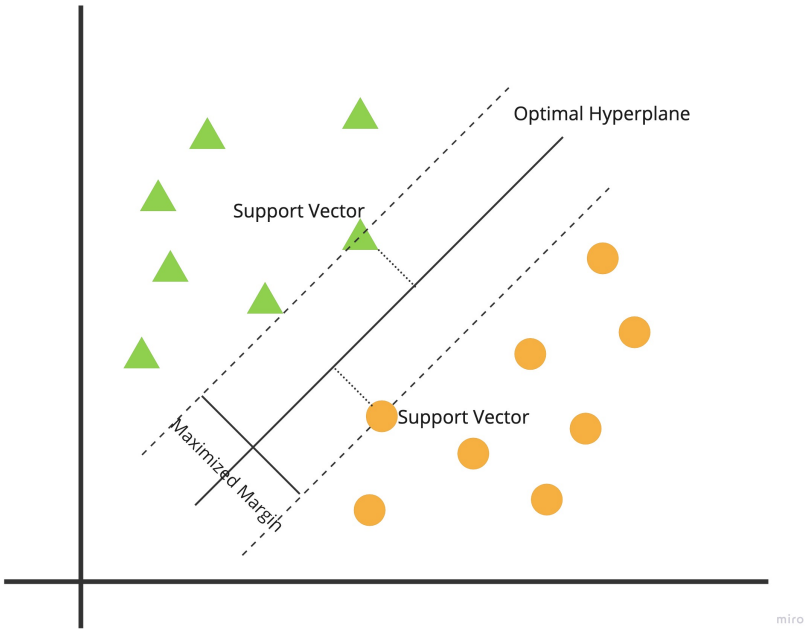
Η απόσταση των x_i από το όριο αποφάσεων, το υπερεπίπεδο δηλαδή, είναι:

$$d_i = \frac{w^T x + \beta}{||w||}$$

Για γραμμικό μοντέλο ταξινόμησης SVM:

$$\hat{y} = \begin{cases} \text{κλάση } 0: & w^T x + \beta < 0 \\ \text{κλάση } 1: & w^T x + \beta \geq 0 \end{cases}$$

Για να πραγματοποιηθεί η κατηγοριοποίηση ο αλγόριθμος προσπαθεί να βρει το βέλτιστο υπερεπίπεδο μέσω της έννοιας του περιθωρίου (margin), στοχεύοντας στη μεγιστοποίησή του. Το υπερεπίπεδο με το μέγιστο περιθώριο ονομάζεται βέλτιστο υπερεπίπεδο (optimal hyperplane).



Σχήμα 3.5: Απόσταση d_i των x_i από το βέλτιστο υπερεπίπεδο

Πλεονεκτήματα των SVM μοντέλων αποτελούν:

- έχουν ισχυρό θεωρητικό υπόβαθρο
- είναι ανθεκτικά στην υπερ-προσαρμογή και λειτουργούν καλά στις γενικεύσεις
- λειτουργούν καλά για πολυδιάστατα σύνολα δεδομένων με μικρό-μεσαίο πλήθος παρατηρήσεων

3.2.2 Μετρικές και Μέθοδοι Αξιολόγησης Μοντέλων Ταξινόμησης

Μη επαρκές πλήθος δεδομένων, μη-αντιπροσωπευτικό δείγμα, ελλιπείς ή ακραίες τιμές αλλά και μεταβλητές που δεν συνεισφέρουν στο αποτέλεσμα, επηρεάζουν σημαντικά τα αποτελέσματα των μοντέλων έτσι ώστε να μην αποτυπώνουν την πραγματική απόδοσή τους. Οι παράγοντες αυτοί μπορεί να οδηγήσουν σε υπερ-προσαρμογή (overfitting) ή υπο-προσαρμογή (underfitting) του συνόλου εκπαίδευσης (training set), απαιτώντας περαιτέρω διενέργειες για τη βελτιστοποίηση των μοντέλων.

Οι μετρικές που εξηγούνται στις επόμενες παραγράφους αποτελούν διαγνωστικούς ελέγχους της λειτουργίας των μοντέλων και βοηθούν στην κατανόηση του τρόπου λειτουργίας τους.

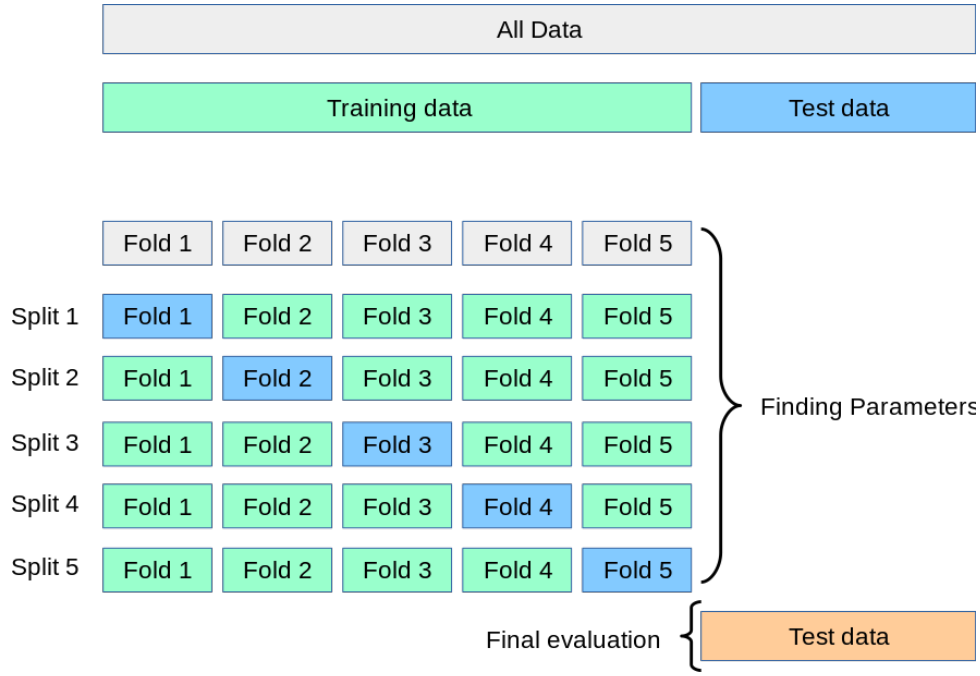
3.2.2.1 Cross Validation

Η μέθοδος Cross Validation είναι από τις πιο γνωστές μεθόδους αξιολόγησης της γενικευμένης ακρίβειας (Accuracy) ενός μοντέλου. Αποτελεί μία μέθοδο επαναδειγματισμού του training set και εκπαίδευσης πολλαπλών μοντέλων ώστε να εκτιμηθεί η ακρίβεια του μοντέλου. Μία γνωστή εκδοχή της είναι η K-Fold Cross Validation, όπου k ορίζεται από τον χρήστη (συνήθως 5 ή 10) και αποτελεί τις επαναλήψεις επαναδειγματισμού.

Βήματα της μεθόδου K-Fold Cross Validation είναι:

1. Χώρισε το σύνολο δεδομένων σε k ίσα μέρη
2. Επίλεξε $k-1$ μέρη ως σύνολο εκπαίδευσης και 1 μέρος ως σύνολο ελέγχου και εκπαίδευσε το μοντέλο
3. Επανάλαβε τα βήματα k φορές, εναλλάσσοντας το σύνολο εκπαίδευσης
4. Υπολόγισε τον μέσο όρο των αποτελεσμάτων

Το Σχήμα 3.6 επεξηγεί τα βήματα του Cross Validation για $k = 5$ επαναλήψεις.



Σχήμα 3.6: K-Fold Cross Validation για $k = 5$

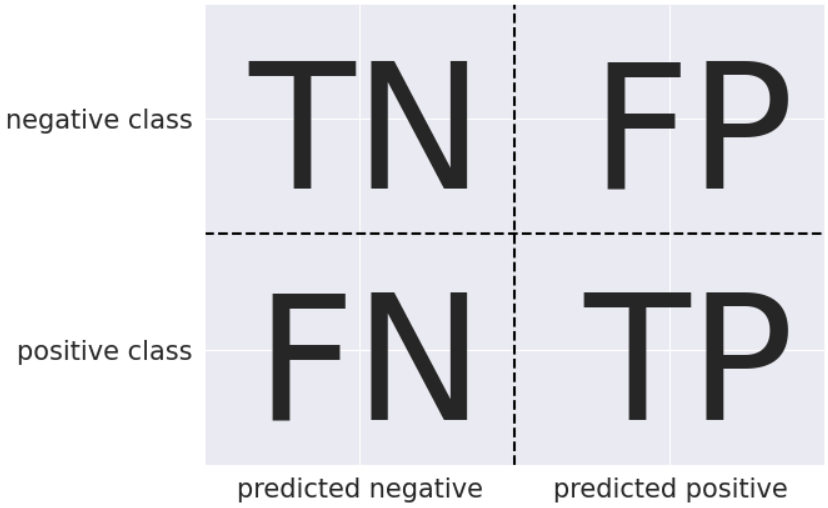
Πηγή: https://scikit-learn.org/stable/modules/cross_validation.html

Εφαρμόζοντας τη μέθοδο κάθε ισομερής διαχωρισμός θα υπάρχει μοναδική φορά στο σύνολο εκπαίδευσης και στο σύνολο ελέγχουtesting set, αξιοποιώντας όλο το σύνολο δεδομένων με πιο αποτελεσματικό τρόπο. Συνδυάζεται με μεθόδους υπερπαραμετροποίησης Hyperparameter Tuning για βελτιστοποίηση του μοντέλου. Μειονέκτημά της μεθόδου είναι το κόστος χρόνου εκτέλεσης που ποικίλει ανάλογα με την πολυπλοκότητα του μοντέλου.

3.2.2.2 Confusion Matrix

Ο Πίνακας Σύγχυσης, Confusion Matrix, χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης. Ο πίνακας συγκρίνει τις εκτιμήσεις του μοντέλου με τις πραγματικές τιμές στις ορισμένες κλάσεις και εκτυπώνει τα αποτελέσματα όπως φαίνεται και από το Σχήμα 3.7.

- True Negative (TN): αριθμός παρατηρήσεων που έχουν ταξινομηθεί ορθά στην κλάση negative class
- False Negative (FN): αριθμός παρατηρήσεων που έχουν ταξινομηθεί λανθασμένα στην κλάση negative class, ενώ στην πραγματικότητα ανήκουν στην positive class
- False Positive (FP): αριθμός παρατηρήσεων που έχουν ταξινομηθεί λανθασμένα στην κλάση positive class, ενώ στην πραγματικότητα ανήκουν στην negative class
- True Positive (TP): αριθμός παρατηρήσεων που έχουν ταξινομηθεί ορθά στην κλάση positive class



Σχήμα 3.7: Confusion Matrix με πραγματικές τιμές στον άξονα y, και εκτιμήσεις στον άξονα x

3.2.2.3 Accuracy

Η μετρική της Ακρίβειας (Accuracy) δίνεται από τη σχέση

$$\text{Accuracy} = \frac{\text{Αριθμός ορθών εκτιμήσεων}}{\text{Συνολικός αριθμός εκτιμήσεων}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.5}$$

3.2.2.4 Precision, Recall & F1-score

Η μετρική Precision υπολογίζει την ακρίβεια των ορθών προβλέψεων που ανήκουν στη θετική κλάση (positive class). Είναι βασική όταν είναι επιθυμητή η μείωση των False Positive (FP) και δίνεται από την εξίσωση:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.6}$$

Χρησιμοποιείται μαζί με τη μετρική Recall που δείχνει πόσες από τις πραγματικά θετικές παρατηρήσεις το μοντέλο μπόρεσε να ταξινομήσει ορθά, και δίνεται από τη σχέση:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.7}$$

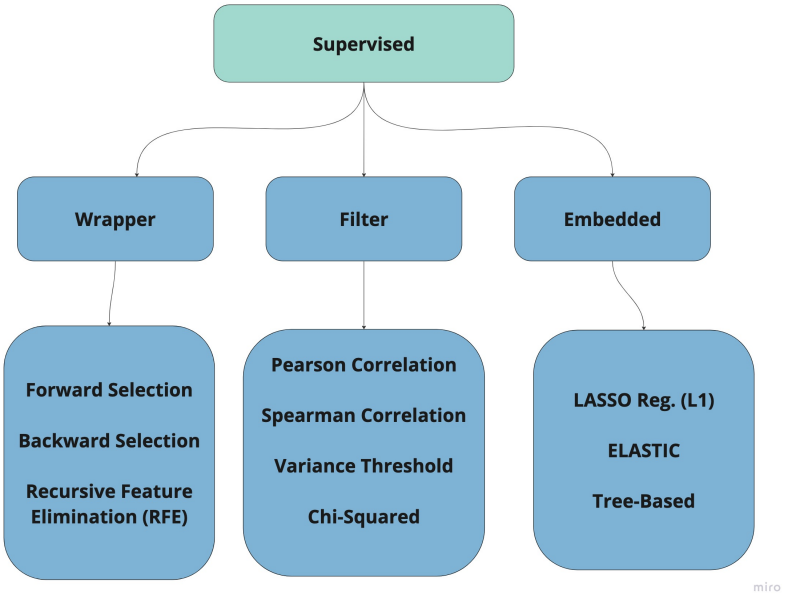
Ο αρμονικός μέσος όρος των μετρικών Recall και Precision ονομάζεται F1-score. Η μετρική αυτή συνδυάζει τις δύο προηγούμενες και είναι ιδιαίτερα βοηθητική στην κατανόηση της απόδοσης του μοντέλου διωνυμικής ταξινόμησης όταν υπάρχει ασυμμετρία παρατηρήσεων (imbalance) μεταξύ των κλάσεων. Ο τύπος του F1-score είναι:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + \frac{FN+FP}{2}} \tag{3.8}$$

3.2.3 Feature Selection και Feature Importance

Η επιλογή των κατάλληλων μεταβλητών (feature selection) αποτελεί βασικό βήματα στη μηχανική μάθηση ως μέθοδος απλοποίησης και μείωσης διάστασης του πίνακα δεδομένων (dimensionality reduction). Κατάλληλη επιλογή features μπορεί να οδηγήσει σε βελτίωση της ακρίβειας των μοντέλων, μείωση χρόνου υπολογισμού, μείωση θορύβου στο σύνολο δεδομένων και καλύτερη ερμηνεία των αποτελεσμάτων.

Υπάρχουν διάφοροι μέθοδοι επιλογής μεταβλητών, τόσο supervised όσο unsupervised. Η παρακάτω εικόνα δείχνει μερικές από τις supervised μεθόδους.



Σχήμα 3.8: Supervised μέθοδοι επιλογής μεταβλητών

- Wrapper μέθοδοι: βασίζονται στον αλγόριθμο που εφαρμόζονται επιλέγοντας υποσύνολα μεταβλητών, αξιολογώντας τα μέσω του Accuracy του μοντέλου, για να προσδιορίσουν τη σημαντικότητα των μεταβλητών. Τείνουν να έχουν υψηλό υπολογιστικό κόστος.
- Filter μέθοδοι: βασίζονται σε στατιστικά χαρακτηριστικά του συνόλου δεδομένων όπως η απόσταση, οι συσχετίσεις κ.λ.π., χωρίς να απαιτεί την εφαρμογή κάποιου μοντέλου.
- Embedded μέθοδοι: συνδυάζουν χαρακτηριστικά των μεθόδων wrapper και filter όμως αποδίδουν χαμηλότερα σε σχέση με τα wrapper.

Recursive Feature Elimination (RFE)

Η μέθοδος Recursive Feature Elimination (RFE), η οποία και εφαρμόζεται στα πλαίσια αυτής της διπλωματικής, αποτελεί μία wrapper μέθοδο διαδοχικής απαλοιφής των λιγότερο σημαντικών μεταβλητών μέχρι να φτάσει στον επιθυμητό αριθμό αυτών που συνεισφέρουν περισσότερο στο μοντέλο.

Μία απλουστευμένη μορφή των βημάτων που ακολουθεί η μέθοδος RFE είναι:

1. Εκπαίδευσε το μοντέλο χρησιμοποιώντας όλα τα features
2. Υπολόγισε τη σημαντικότητα κάθε feature βάσει των συντελεστών (coefficient attribute) τους ή βάσει του βαθμού κατάταξης αυτών (feature importance attribute) ανάλογα το μοντέλο
3. Αφαίρεσε τα n λιγότερο σημαντικά features έως ότου μείνουν τα k σημαντικότερα
4. Εκπαίδευσε το μοντέλο χρησιμοποιώντας τα k σημαντικότερα features

Feature Importance

Feature Importance, δηλαδή η σημαντικότητα των μεταβλητών, δείχνει το πόσο συνεισφέρει κάθε μεταβλητή στις προβλέψεις που κάνει ένα μοντέλο και στην εν γένει απόδοσή του. Ο υπολογισμός του Feature Importance μπορεί να επιτευχθεί με διάφορους τρόπους. Για παράδειγμα, για γραμμικά μοντέλα όπως Logistic Regression, SVM ή Linear Regression, υπολογίζονται οι συντελεστές των μεταβλητών, ενώ για μοντέλα βασιζόμενα σε δέντρα αποφάσεων, όπως είναι Random Forest, Gradient Boosting, XGBoost, ο υπολογισμός γίνεται βάσει του διαχωρισμού των κόμβων.

Συμπερίληψη μη σημαντικών μεταβλητών ή υποτίμηση σχετικών μεταβλητών μπορεί να οδηγήσει σε χαμηλή απόδοση ενός μοντέλου, για τον λόγο αυτό οι προαναφερθείσες μέθοδοι επιλογής μεταβλητών και η ερμηνεία τους αποτελεί βασικό στάδιο στην κατανόηση εφαρμογής ενός μοντέλου μηχανικής μάθησης. Επομένως, το Feature Importance αποτελεί ένα χρήσιμο model agnostic στα πλαίσια της έννοιας του eXplainable Artificial Intelligence (XAI), αρχές του οποίου αποτελούν η επεξηγηματικότητα, η γνώση των περιορισμών, η ερμηνεία των αποτελεσμάτων και της απόδοσης με όρους κατανοητούς για όλους.

Κεφάλαιο 4

Επεξεργασία και καθαρισμός δεδομένων

Στο παρόν κεφάλαιο θα γίνει η ανάλυση των δεδομένων που συλλέχθηκαν από το Matomo για το e-Service της Πολωνίας.

Η άντληση των ανεπεξέργαστων δεδομένων από το Matomo, μορφής csv, αφορούν:

- τις βασικές μετρικές του συστήματος (Main Metrics)
- τα κανάλια εισόδου (Channel Type)

Η επεξεργασία και ανάλυση των δεδομένων πραγματοποιήθηκε με τη χρήση της γλώσσας προγραμματισμού Python.

4.1 Μετατροπή Τύπων Μεταβλητών

Είναι πιθανό σε μεγάλους όγκους δεδομένων οι τύποι ορισμένων μεταβλητών να μην βρίσκονται στην επιθυμητή μορφή για περαιτέρω υπολογισμούς ή επεξεργασία δυσχεραίνοντας το έργο του αναλυτή. Παραδείγματος χάριν, χρονικές μεταβλητές ή μεταβλητές που περιέχουν παρατηρήσεις μονάδων μέτρησης μπορεί να έχουν αποθηκευτεί σε μορφή συμβολοσειρών (**string**). Γενικά, ενδέχεται να υπάρχουν μεταβλητές που έχουν αποθηκευτεί σε τύπο διαφορετικό από τον επιθυμητό, επομένως, είναι σημαντικό να ελεγχθούν και να κατανοηθούν οι τύποι των μεταβλητών, αλλά και να γίνουν οι απαραίτητες ενέργειες για την μετατροπή τους.

Ο έλεγχος αυτός πραγματοποιείται εύκολα με την εντολή `pandas.DataFrame.info()` που δίνει ως αποτέλεσμα συνοπτικές πληροφορίες του πίνακα προς μελέτη όπως:

- αριθμό στηλών και γραμμών

- μεταβλητές
- αριθμό παρατηρήσεων ανά μεταβλητή
- τύπος κάθε μεταβλητής

Στην εν λόγω περίπτωση, η ανάγκη για αυτή την επεξεργασία και των βημάτων που αναλύονται παρακάτω, προέκυψε μόνο για τον πίνακα Main Metrics, ενώ για τον πίνακα Channel Type δεν απαιτήθηκε καμία τέτοια ενέργεια.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 37 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Date                                        504 non-null    datetime64[ns]
 1   Visitors from Search Engines                504 non-null    int64
 2   Visitors from Social Networks               504 non-null    int64
 3   Visitors from Direct Entry                  504 non-null    int64
 4   Visitors from Websites                      504 non-null    int64
 5   Visitors from Campaigns                     504 non-null    int64
 6   Bounce Rate                                 504 non-null    object
 7   Actions per Visit                           504 non-null    float64
 8   Avg. Visit Duration (in seconds)            504 non-null    object
 9   Avg. Duration of a Returning Visit (in sec) 504 non-null    object
 10  Avg. Actions per Returning Visit            504 non-null    float64
 11  Bounce Rate for Returning Visits            504 non-null    object
 12  Avg. Duration of a New Visit (in sec)       504 non-null    object
 13  Avg. Actions per New Visit                  504 non-null    float64
 14  Bounce Rate for New Visits                  504 non-null    object
 15  Unique visitors                             501 non-null    float64
 16  Visits                                      501 non-null    float64
 17  Actions                                     501 non-null    float64
 18  Maximum actions in one visit                501 non-null    float64
 19  Bounces                                     501 non-null    float64
 20  Total time spent by visitors (in seconds)   501 non-null    float64
 21  New Visits                                  482 non-null    float64
 22  Actions by New Visits                       482 non-null    float64
 23  Unique new visitors                         482 non-null    float64
 24  max_actions_new                             482 non-null    float64
 25  Pageviews                                   501 non-null    float64
 26  Unique Pageviews                            501 non-null    float64
 27  Downloads                                   501 non-null    float64
 28  Unique Downloads                            501 non-null    float64
 29  Outlinks                                    501 non-null    float64
 30  Unique Outlinks                             501 non-null    float64
 31  Avg. generation time                        504 non-null    object
 32  Returning Visits                            491 non-null    float64
 33  Actions by Returning Visits                 491 non-null    float64
 34  Unique returning visitors                   491 non-null    float64
 35  Maximum actions in one returning visit      491 non-null    float64
 36  Distinct websites                           473 non-null    float64
dtypes: datetime64[ns](1), float64(24), int64(5), object(7)
memory usage: 145.8+ KB
```

Σχήμα 4.1: Πληροφορίες Polish Main Metrics

Στον παραπάνω Πίνακα του Σχήματος 4.1 φαίνονται οι συνοπτικές πληροφορίες του πίνακα Main Metrics του πολωνικού στιγμιοτύπου αμέσως μετά την διαγραφή των περιττών στηλών και χωρίς άλλη επεξεργασία. Φαίνεται πως οι μεταβλητές Avg. Visit Duration (in seconds), Avg. Duration of a Returning visit (in sec), Avg. Duration of a New Visit (in sec) και Avg. generation time ενώ αναφέρονται σε χρονικές μετρήσεις, έχουν αποθηκευτεί ως τύπος **object**, με διαφορετικό μήκος (length) για κάθε παρατήρηση.

Αντίστοιχα, παρατηρούμε πως το ίδιο συμβαίνει και για τις μεταβλητές Bounce Rate, Bounce Rate for Returning Visits, Bounce Rate for New Visits που οι παρατηρήσεις τους, αν και εκφράζουν ποσοστό, είναι αποθηκευμένες ως τύπος **object** με μορφή x%.

Ακολουθούνται οι εξής διαδικασίες:.

- **Μετατροπή των χρονικών μεταβλητών string σε float**.

  Με την εντολή `pd.to_timedelta` μετατρέπονται οι τύποι string σε float για τις μετρικές χρόνου και στη συνέχεια τα δευτερόλεπτα μετατρέπονται σε λεπτά. Παρατίθεται ενδεικτικός κώδικας.

```
pol['Avg.Visit_Duration (in min)'] = pd.to_timedelta(pol['Avg. Visit
↪  Duration (in seconds)'])
pol['Avg.Visit_Duration (in min)'] = pol['Avg.Visit_Duration (in
↪  min)'].dt.total_seconds()/60
```

  Προκύπτουν νέες στήλες λεπτών τύπου float και διαγράφονται οι στήλες τύπου string.

	Total time spent by visitors (in min)	Total time spent by visitors (in seconds)	Avg.Visit_Duration (in min)	Avg. Visit Duration (in seconds)	Avg.Duration_New Visit (in min)	Avg. Duration of a New Visit (in sec)
0	41.50	2,490.00	4.62	4 min 37s	2.87	2 min 52s
1	1.63	98.00	0.42	25s	0.42	25s
2	13.03	782.00	6.52	6 min 31s	0.00	0s
3	0.00	0.00	0.00	0s	0.00	0s
4	41.18	2,471.00	13.73	13 min 44s	16.87	16 min 52s
...	...	...	...	...	...	...
499	632.27	37,936.00	2.82	2 min 49s	2.02	2 min 1s
500	1,802.32	108,139.00	3.35	3 min 21s	0.92	55s
501	1,270.22	76,213.00	2.50	2 min 30s	1.73	1 min 44s
502	1,642.22	98,533.00	3.80	3 min 48s	1.95	1 min 57s
503	1,478.40	88,704.00	3.35	3 min 21s	2.03	2 min 2s

504 rows × 6 columns

Σχήμα 4.2: Polish - Χρονικές μεταβλητές στον αρχικό και τελικό τους τύπο

- **Μετατροπή ποσοστιαίων string σε int και αφαίρεση συμβόλου %**

  Σε αυτή την περίπτωση, η διαδικασία είναι ευκολότερη, καθώς οι μόνες ενέργειες που απαιτούνται είναι η αφαίρεση του συμβόλου % από το τέλος του κάθε string και, εν συνεχεία, η μετατροπή αυτού σε ακέραιο. Η διαδικασία είναι η ίδια για κάθε μία από τις στήλες Bounce Rate, Bounce Rate for Returning Visits, Bounce Rate for New Visits. Όπως και στην περίπτωση των χρονικών μεταβλητών, οι μετατροπές αποθηκεύονται σε νέες στήλες.

```
# Removing the % symbol from each column
# and saving the result in a new column
pol['Bounce Rate(%)'] = list(map(lambda x: x[:-1], pol['Bounce Rate'].values))

pol['Bounce Rate New Visits(%)'] = list(map(lambda x: x[:-1], pol['Bounce Rate for New Visits'].values))

pol['Bounce Rate Returning Visits(%)'] = list(map(lambda x: x[:-1], pol['Bounce Rate for Returning Visits'].values))

# Converting each new column into int
pol['Bounce Rate(%)'] = [int(x) for x in pol['Bounce Rate(%)'].values]

pol['Bounce Rate New Visits(%)'] = [int(x) for x in pol['Bounce Rate New Visits(%)'].values]

pol['Bounce Rate Returning Visits(%)'] = [int(x) for x in pol['Bounce Rate Returning Visits(%)'].values]
```

	Bounce Rate	Bounce Rate(%)	Bounce Rate for New Visits	Bounce Rate New Visits(%)	Bounce Rate for Returning Visits	Bounce Rate Returning Visits(%)
0	22%	22	25%	25	20%	20
1	25%	25	25%	25	0%	0
2	50%	50	0%	0	50%	50
3	100%	100	100%	100	0%	0
4	33%	33	50%	50	0%	0
...	...	...	...	...	...	...
499	61%	61	67%	67	57%	57
500	64%	64	76%	76	58%	58
501	63%	63	72%	72	58%	58
502	60%	60	72%	72	52%	52
503	59%	59	69%	69	55%	55

504 rows × 6 columns

Σχήμα 4.3: Polish - Ποσοτικές μεταβλητές στον αρχικό και τελικό τους τύπο

Με την ολοκλήρωση των παραπάνω βημάτων, ο αρχικός πίνακας (Data Frame) έχει καθαριστεί από τύπους μεταβλητών που δεν μπορούσαν να χρησιμοποιηθούν για υπολογισμούς και στατιστικά μοντέλα. Οι αρχικές μεταβλητές μη επιθυμητών τύπων διαγράφονται για αποφυγή σύγχυσης, και ο πίνακας που προκύπτει τελικά είναι:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 36 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Date                                    504 non-null    datetime64[ns]
 1   Visitors from Search Engines            504 non-null    int64
 2   Visitors from Social Networks           504 non-null    int64
 3   Visitors from Direct Entry              504 non-null    int64
 4   Visitors from Websites                  504 non-null    int64
 5   Actions per Visit                       504 non-null    float64
 6   Avg. Actions per Returning Visit        504 non-null    float64
 7   Avg. Actions per New Visit              504 non-null    float64
 8   Unique visitors                         501 non-null    float64
 9   Visits                                  501 non-null    float64
 10  Actions                                 501 non-null    float64
 11  Maximum actions in one visit            501 non-null    float64
 12  Bounces                                 501 non-null    float64
 13  New Visits                              482 non-null    float64
 14  Actions by New Visits                   482 non-null    float64
 15  Unique new visitors                     482 non-null    float64
 16  max_actions_new                         482 non-null    float64
 17  Pageviews                               501 non-null    float64
 18  Unique Pageviews                        501 non-null    float64
 19  Downloads                               501 non-null    float64
 20  Unique Downloads                        501 non-null    float64
 21  Outlinks                                501 non-null    float64
 22  Unique Outlinks                         501 non-null    float64
 23  Returning Visits                        491 non-null    float64
 24  Actions by Returning Visits             491 non-null    float64
 25  Unique returning visitors               491 non-null    float64
 26  Maximum actions in one returning visit  491 non-null    float64
 27  Distinct websites                       473 non-null    float64
 28  Bounce Rate(%)                          504 non-null    int64
 29  Bounce Rate New Visits(%)               504 non-null    int64
 30  Bounce Rate Returning Visits(%)         504 non-null    int64
 31  Avg.Visit_Duration (in min)             504 non-null    float64
 32  Avg.Duration_Returning Visit (in min)   504 non-null    float64
 33  Avg.Duration_New Visit (in min)         504 non-null    float64
 34  Avg.Generation Time (in min)            504 non-null    float64
 35  Total time spent by visitors (in min)   501 non-null    float64
dtypes: datetime64[ns](1), float64(28), int64(7)
memory usage: 141.9 KB
```

Σχήμα 4.4: Main Metrics - Μετά τον μετασχηματισμό τύπων

4.2 Εκτίμηση Ελλιπών Παρατηρήσεων με kNN Imputation

Στην πορεία διερεύνησης και κατανόησης των δεδομένων των Data Frame που δημιουργήθηκαν για την μελέτη των παραπάνω διαστημάτων, εντοπίστηκαν ελλιπείς παρατηρήσεις σε ορισμένες μεταβλητές. Λόγω του μικρού πλήθους παρατηρήσεων επιλέχθηκε η μέθοδος kNN Imputation, εισάγοντας μικρή μόνο μεροληψία, για να μην χαθεί επιπλέον πληροφορία. Για τον έλεγχο των μεταβλητών που περιέχουν τέτοιες παρατηρήσεις, του πλήθους αυτών, καθώς και την ύπαρξη τυχόν μοτίβου στον τρόπο που λείπουν, χρησιμοποιήθηκαν τρία διαφορετικά εργαλεία της Python και για τα τρία στιγμιότυπα.

- `pandas.DataFrame.info()`

  Εμφανίζει συνοπτικές πληροφορίες για τη δομή του Data Frame, τον τύπο, τον αριθμό των μεταβλητών καθώς και των μη μηδενικών (non-null) τιμών τους.

- `pandas.DataFrame.isnull().sum()`

  Συμπληρωματική της προηγούμενης εντολής, η εντολή `pandas.Data.Frame.isna().sum()` εμφανίζει το σύνολο των ελλιπών παρατηρήσεων για κάθε μεταβλητή του Data Frame. Ενώ η εντολή `df['Column Name'].index[df['Column Name'].apply(np.isnan)]` επιστρέφει την θέση κάθε παρατήρησης που λείπει για δοσμένη μεταβλητή (στήλη πίνακα).

- Missingno βιβλιοθήκη

  Η βιβλιοθήκη παρέχει τρία γραφήματα (matrix, heatmap, bar) για την γρήγορη απεικόνιση των δεδομένων ενός Data Frame, την κατανόηση της δομής των δεδομένων, της συσχέτισης μεταξύ των στηλών καθώς και τον αριθμό των ελλιπών παρατηρήσεων. Με την απεικόνιση αυτή, επιβεβαιώθηκε γρήγορα η ύπαρξη ή μη, ελλιπών παρατηρήσεων στους πίνακες των τριών στιγμιοτύπων και εξετάστηκε εάν οι ελλιπείς παρατηρήσεις ακολουθούν κάποιο μοτίβο.

- `pandas.DataFrame.drop()`

  Στήλες με έλλειψη 60% και πάνω των δεδομένων τους αυτομάτως απαλείφονται καθώς δεν μπορούν να συνεισφέρουν στην ανάλυση. Η απαλοιφή τους πραγματοποιείται είτε χειροκίνητα στο αρχείο csv, είτε μέσω της εντολής `pandas.DataFrame.drop()`. Συγκεκριμένα 28/65 στήλες χρειάστηκε να διαγραφούν με αυτόν τον τρόπο λόγω ελλιπών ή μηδενικών παρατηρήσεων.

- `sklearn.impute.KNNImputer`

  Εφαρμόζει την μέθοδο kNN Imputation για την συμπλήρωση των ελλιπών τιμών βάσει των k-πλησιέστερων γειτόνων. Στην προκειμένη περίπτωση χρησιμοποιήθηκαν $k = 10$ γείτονες βάσει ελέγχου διαφορετικών τιμών.

Οι παραπάνω εντολές εφαρμόζονται στους πίνακες του πολωνικού e-Service.

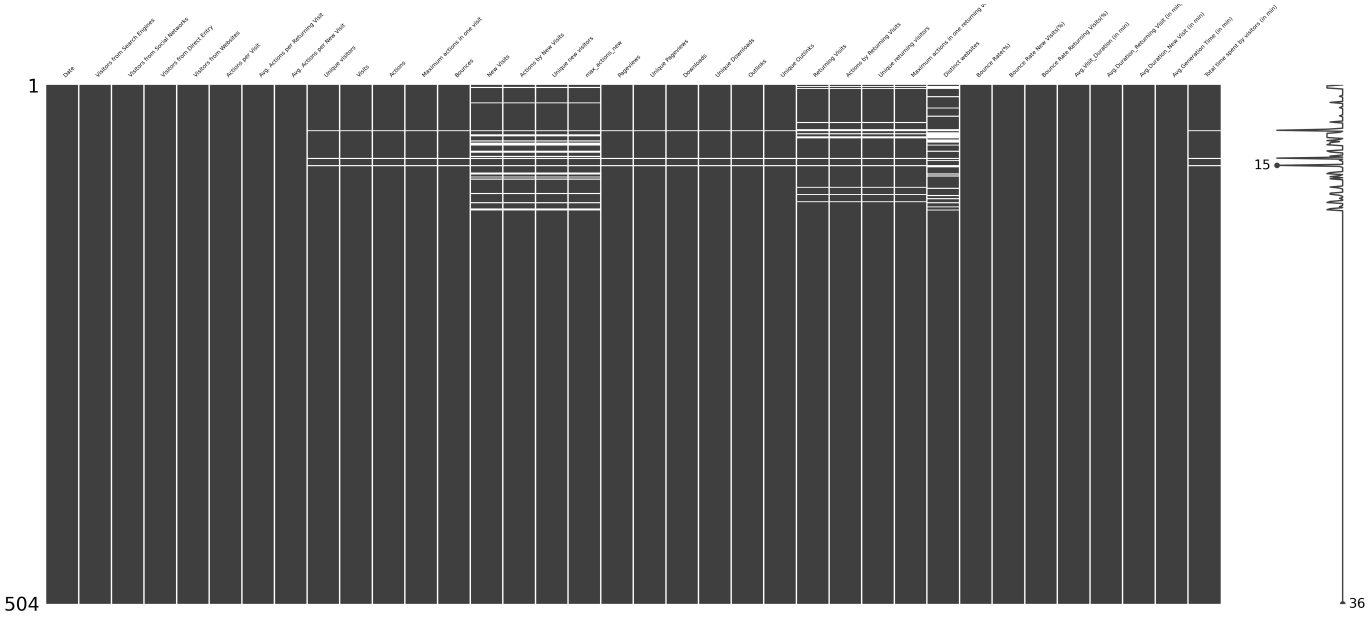
```
Date                                          0
Visitors from Search Engines                  0
Visitors from Social Networks                 0
Visitors from Direct Entry                    0
Visitors from Websites                        0
Actions per Visit                             0
Avg. Actions per Returning Visit              0
Avg. Actions per New Visit                    0
Unique visitors                               3
Visits                                        3
Actions                                       3
Maximum actions in one visit                  3
Bounces                                       3
New Visits                                   22
Actions by New Visits                        22
Unique new visitors                          22
max_actions_new                              22
Pageviews                                     3
Unique Pageviews                              3
Downloads                                     3
Unique Downloads                              3
Outlinks                                      3
Unique Outlinks                               3
Returning Visits                             13
Actions by Returning Visits                  13
Unique returning visitors                    13
Maximum actions in one returning visit       13
Distinct websites                            31
Bounce Rate(%)                                0
Bounce Rate New Visits(%)                     0
Bounce Rate Returning Visits(%)               0
Avg.Visit_Duration (in min)                   0
Avg.Duration_Returning Visit (in min)         0
Avg.Duration_New Visit (in min)               0
Avg.Generation Time (in min)                  0
Total time spent by visitors (in min)         3
dtype: int64
```

Σχήμα 4.5: Main Metrics - Πλήθος ελλιπών παρατηρήσεων για τις εναπομείνασες 36 μεταβλητές μετά την αφαίρεση στηλών

```
Date                                0
Label                               0
Unique visitors                     0
Visits                              0
Actions                             0
Maximum actions in one visit        0
Bounces                             0
Total time by visitors (min)        0
dtype: int64
```

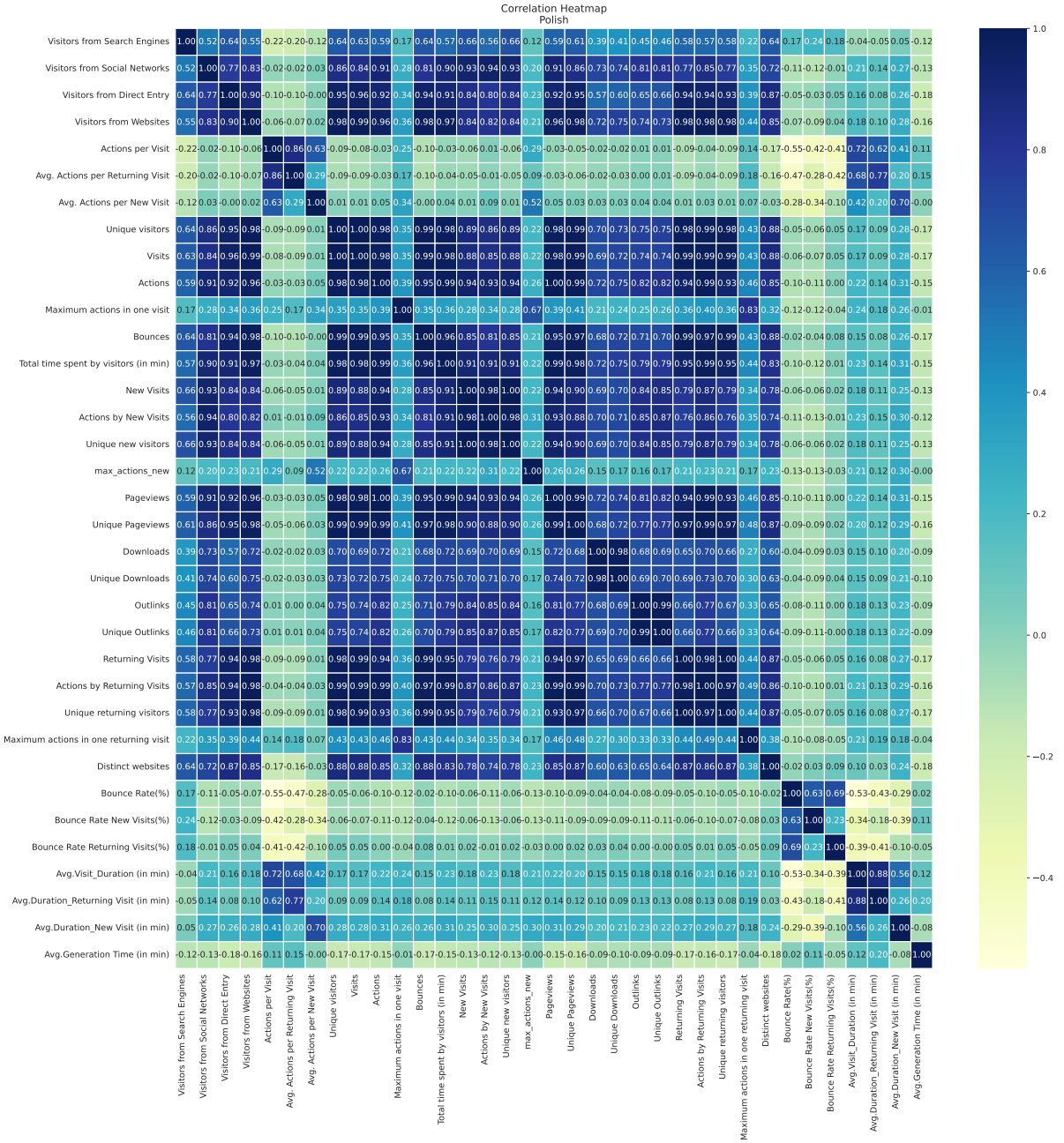
Σχήμα 4.6: Channel - Πλήθος ελλιπών παρατηρήσεων

Παρατηρείται πως στον πίνακα Channel, Σχήμα 4.6, δεν υπάρχουν ελλιπείς τιμές, επομένως δεν απαιτείται περαιτέρω ενέργεια. Αντιθέτως, στον πίνακα των Main Metrics βλέπουμε πως 21/36 μεταβλητές περιέχουν άγνωστες παρατηρήσεις, ωστόσο ο αριθμός τους είναι αναλογικά μικρός σε σχέση με το συνολικό μέγεθος των παρατηρήσεων (n = 504), αλλά όχι αμελητέος.



Σχήμα 4.7: Missingno Matrix για Main Metrics

Το σχήμα 4.7 δείχνει τα μοτίβα των ελλιπών τιμών και τις θέσεις τους. Φαίνεται να υπάρχει κάποιο μοτίβο συσχέτισης έλλειψης τιμών μεταξύ ορισμένων μεταβλητών, π.χ. όπου η New Visits εμφανίζει ελλιπείς τιμές, εμφανίζει και οι μεταβλητή Action by New Visits, Unique new visitors, max_actions_new. Αντίστοιχα για τις Returning Visits, Actions by Returning Visits, Unique returning visitors και Maximum actions in one returning visit. Η συσχέτιση επιβεβαιώνεται με τον πίνακα Heatmap, Σχήμα 4.8 που δείχνει τον βαθμό συσχέτισης μεταξύ των μεταβλητών (μπλε - θετική συσχέτιση, κίτρινο - αρνητική). Επομένως, υπάρχει συσχέτιση μεταξύ των προαναφερθέντων μεταβλητών ανά ομάδες.



Σχήμα 4.8: Missingno Heatmap για Main Metrics

Προκύπτει πως ο μηχανισμός που ακολουθούν οι ελλιπείς τιμές του Data Frame Main Metrics μπορεί να χαρακτηριστεί ως MAR, καθώς υπάρχει κάποια συσχέτιση μεταξύ των μεταβλητών, όμως δεν είναι γνωστό εάν ο μηχανισμός σχετίζεται με την τιμή τους. Για τον λόγο αυτό, εφαρμόζεται η kNN Imputation με k=10 γείτονες. Ο αλγόριθμος αντικαθιστά τις "NaN" (Not a Number) και επιστρέφει πίνακα με τις υπολογισμένες τιμές. Παρατίθενται ενδεικτικά μερικές γραμμές του πίνακα με τις άγνωστες τιμές να διαμορφώνονται μετά την χρήση της kNN Imputation.

	Date	Visits	New Visits	Total time spent by visitors (in min)	Actions by New Visits	Pageviews
40	2019-12-18	10.0	4.0	133.2	28.0	83.0
41	2019-12-19	11.0	4.0	67.7	32.0	60.0
42	2019-12-20	10.0	5.0	45.7	9.0	33.0
43	2019-12-21	1.0	1.0	0.0	1.0	1.0
44	2019-12-22	nan	nan	nan	nan	nan
45	2019-12-23	7.0	2.0	2.3	10.0	16.0
46	2019-12-24	17.0	16.0	1.0	30.0	32.0
47	2019-12-25	4.0	4.0	0.1	6.0	6.0
48	2019-12-26	1.0	nan	0.0	nan	1.0
49	2019-12-27	2.0	nan	12.6	nan	4.0
50	2019-12-28	4.0	4.0	0.0	5.0	5.0
51	2019-12-29	1.0	1.0	0.0	1.0	1.0
52	2019-12-30	3.0	1.0	0.0	1.0	4.0
53	2019-12-31	15.0	13.0	71.8	46.0	48.0
54	2020-01-01	1.0	nan	0.0	nan	1.0
55	2020-01-02	7.0	2.0	93.7	86.0	92.0
56	2020-01-03	5.0	nan	46.7	nan	60.0
57	2020-01-04	2.0	nan	3.5	nan	11.0
58	2020-01-05	1.0	nan	0.0	nan	2.0
59	2020-01-06	2.0	1.0	16.5	1.0	17.0

(α′) Ελλιπείς τιμές

	Date	Visits	New Visits	Total time spent by visitors (in min)	Actions by New Visits	Pageviews
40	2019-12-18	10.0	4.0	133.2	28.0	83.0
41	2019-12-19	11.0	4.0	67.7	32.0	60.0
42	2019-12-20	10.0	5.0	45.7	9.0	33.0
43	2019-12-21	1.0	1.0	0.0	1.0	1.0
44	2019-12-22	5.7	2.7	54.8	17.4	42.8
45	2019-12-23	7.0	2.0	2.3	10.0	16.0
46	2019-12-24	17.0	16.0	1.0	30.0	32.0
47	2019-12-25	4.0	4.0	0.1	6.0	6.0
48	2019-12-26	1.0	2.3	0.0	7.4	1.0
49	2019-12-27	2.0	2.0	12.6	6.4	4.0
50	2019-12-28	4.0	4.0	0.0	5.0	5.0
51	2019-12-29	1.0	1.0	0.0	1.0	1.0
52	2019-12-30	3.0	1.0	0.0	1.0	4.0
53	2019-12-31	15.0	13.0	71.8	46.0	48.0
54	2020-01-01	1.0	2.3	0.0	7.4	1.0
55	2020-01-02	7.0	2.0	93.7	86.0	92.0
56	2020-01-03	5.0	4.1	46.7	24.8	60.0
57	2020-01-04	2.0	2.0	3.5	7.9	11.0
58	2020-01-05	1.0	2.0	0.0	6.4	2.0
59	2020-01-06	2.0	1.0	16.5	1.0	17.0

(β′) kNN εκτιμώμενες τιμές

Σχήμα 4.9: (Αριστερά) σειρές με ελλιπείς τιμές - (Δεξιά) εκτιμώμενες τιμές μετά την εφαρμογή kNN Imputation

Η πληρότητα πλέον του πίνακα επιβεβαιώνεται και από την εντολή `pandas.DataFrame.isnull().sum()`. Έχοντας όλες τις τιμές στον πίνακα των Main Metrics, είναι δυνατό να πραγματοποιηθεί η ανάλυση και η εφαρμογή μοντέλων μηχανικής μάθησης.

Μετά την εφαρμογή του καθαρισμού των δεδομένων το Data Frame των Main Metrics διαμορφώνεται ως εξής:

Όνομα Data Frame	pol
Αριθμός Μεταβλητών	65
Αριθμός Παρατηρήσεων	504
Τύποι Μεταβλητών	datetime float int object(string)
Ελλιπείς Τιμές	NaN Null μεταβλητές με >60% NaN ή Null

Όνομα Data Frame	polish
Αριθμός Μεταβλητών	36
Αριθμός Παρατηρήσεων	504
Τύποι Μεταβλητών	datetime float int
Ελλιπείς Τιμές	-

Πίνακας 4.1: Αρχικό και τελικό στάδιο Data Frame Main Metrics

4.3 Επιλογή Χρονικών Περιόδων

Για το πολωνικό e-Service η πρώτη καταχώρηση είναι 8 Νοεμβρίου 2019 και η τελευταία 25 Μαρτίου 2021. Το διάστημα αυτό περιλαμβάνει την πανδημία Covid-19, πρώτο και δεύτερο κύμα έξαρσης, και συνεπώς, την υποχρεωτική τηλεκπαίδευση που επηρέασε την χρήση της πλατφόρμας.

Στη μελέτη γίνεται χρήση όλων των τιμών, ωστόσο στην περιγραφική και διερευνητική ανάλυση γίνεται τμηματοποίηση των δεδομένων βάσει των κυμάτων της πανδημίας για στοχευμένη κατανόηση της επιρροής αυτών στην πλατφόρμα.

Διαστήματα Ανάλυσης	
Pre Covid-19	8 Νοεμβρίου 2019 - 9 Μαρτίου 2020
Wave 1	10 Μαρτίου 2020 - 24 Μαΐου 2020
Wave 2	19 Οκτωβρίου 2020 - 17 Ιανουαρίου 2021
Post Covid-19	18 Ιανουαρίου 2021 - 25 Μαρτίου 2021

Πίνακας 4.2: Διαστήματα διερευνητικής ανάλυσης βάσει κυμάτων πανδημίας

Από την προηγούμενη Παράγραφο 4.2 εύρεσης και εκτίμησης ελλιπών τιμών προέκυψε πως αυτές εντοπίζονται μόνο στην περίοδο Pre Covid-19, κι αυτό διότι το Up2U χρησιμοποιούταν ακόμη πιλοτικά και στα πλαίσια εκπαίδευσης των δασκάλων κατά την περίοδο αυτή. Η ευρεία χρήση της συστηματοποιήθηκε σημαντικά από την πανδημία και τον πρώτο εγκλεισμό.

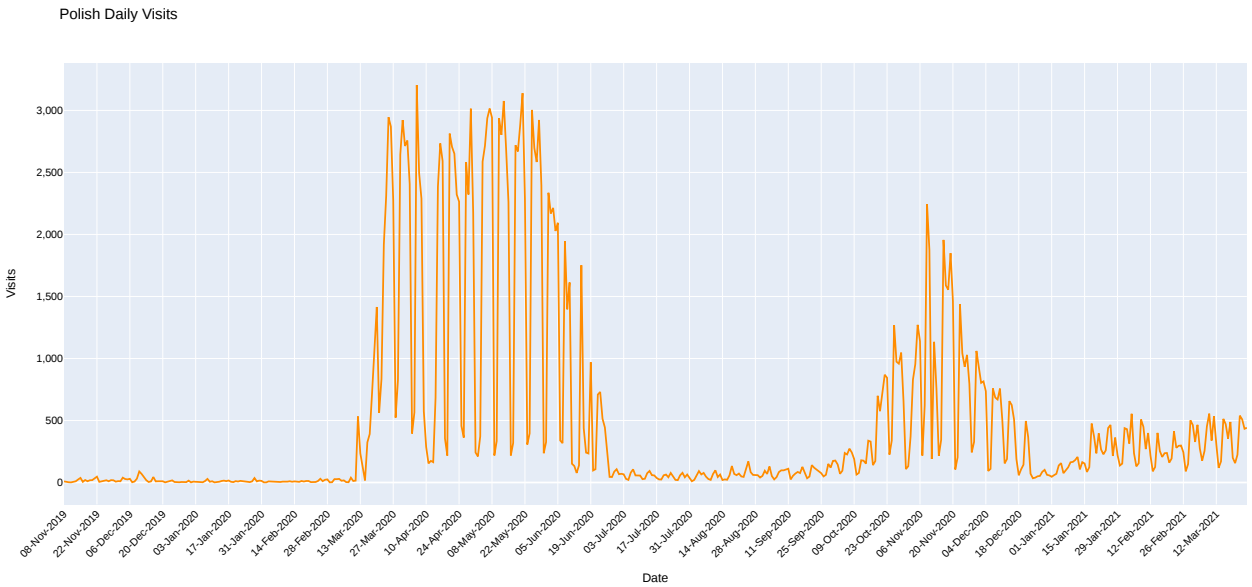
Αξίζει να σημειωθεί ακόμη πως τα διαστήματα δεν αποτελούνται από το ίδιο πλήθος δεδομένων, ωστόσο επιλέγεται αυτή η τμηματοποίηση για αναλογική σύγκριση και την καλύτερη δυνατή κατανόηση της χρήσης της πλατφόρμας στο σύνολό της αλλά και ανά περίοδο. Στο Κεφάλαιο 5 που περιλαμβάνει την εφαρμογή των μοντέλων μηχανικής μάθησης αξιοποιείται όλο το Data Frame.

Κεφάλαιο 5

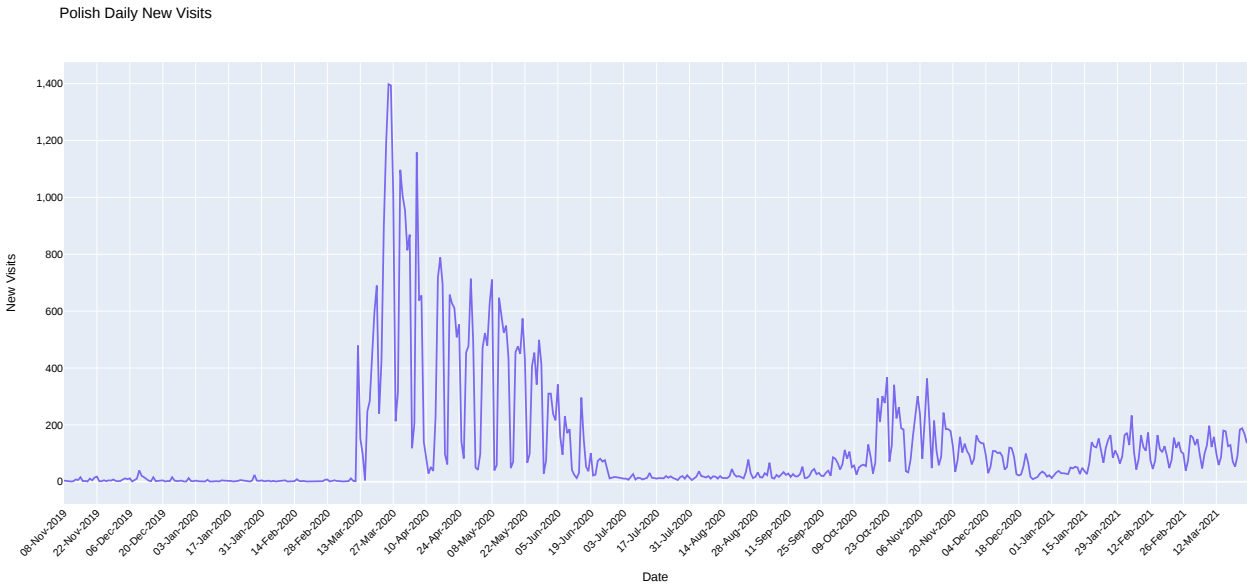
Περιγραφική Ανάλυση και Εφαρμογή Μοντέλων Μηχανικής Μάθησης

5.1 Περιγραφική & Διερευνητική Ανάλυση

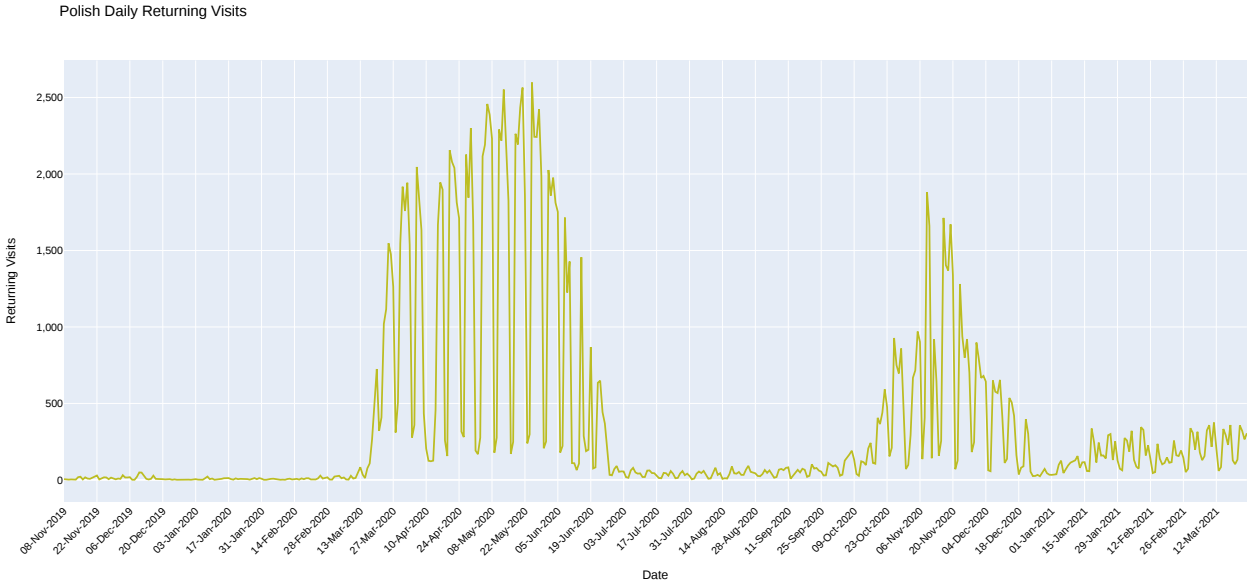
Σκοπός της παραγράφου είναι η ανάλυση μετρικών σχετικών με τον τύπο επισκεπτών (Visitor Type Metrics), δραστηριοτήτων στην πλατφόρμα (Activity Metrics), και του Bounce Rate(%). Όπως προαναφέρθηκε, η διερευνητική ανάλυση γίνεται βάσει των κυμάτων της πανδημίας Covid-19 στην Πολωνία, όπως αναγράφονται στον Πίνακα 4.2. Η επιλογή των διαστημάτων κατανοείται καλύτερα παρατηρώντας τα ακόλουθα γραφήματα ημερήσιων επισκέψεων.



Σχήμα 5.1: Ημερήσια επισκεψιμότητα πολωνικού e-Service



Σχήμα 5.2: Ημερήσια επισκεψιμότητα νέων χρηστών



Σχήμα 5.3: Ημερήσια επισκεψιμότητα χρηστών που επιστρέφουν

Από τα παραπάνω γραφήματα προκύπτουν τα εξής:

- περιορισμένη χρήση της πλατφόρμας μέχρι τον Μάρτιο του 2020 όπου εμφανίζεται κατακόρυφη αύξηση σε όλα τα είδη επισκεπτών
- κορύφωση της χρήσης με συστηματικότητα την περίοδο Μαρτίου - Ιουνίου, όπως φαίνεται από τα ημερήσια γράφημα των Visits και Returning Visits
- νέα περίοδος ραγδαίας αύξησης τον Οκτωβρίου του 2020, σε μικρότερο βαθμό από την περίοδο κορύφωση
- μείωση τον Ιανουάριο του 2021 ωστόσο η υπηρεσία διατηρεί συστηματικότητα στη χρήση της και αυξημένα επίπεδα σε σχέση με την περίοδο του Νοεμβρίου - Απριλίου
- υπάρχει σχετική ομοιότητα στον τρόπο που καταγράφονται οι μεταβλητές Visits και Returning Visits που σημαίνει πως αρκετοί από τους ήδη υπάρχοντες χρήστες επισκέπτονταν την υπηρεσία. Αντιθέτως η μεταβλητή New Visits κορυφώνεται στις 25 Μαρτίου και μετά εμφανίζει μειωμένα επίπεδα κορυφών, που ενδεχομένως να σημαίνει πως η πλατφόρμα έφτασε στη μέγιστη χρήση της την περίοδο του Μαρτίου 2020

5.1.1 Περίοδος Pre Covid-19

Η περίοδος Pre Covid-19 αναφέρεται στις ημερομηνίες 8 Νοεμβρίου 2019 - 9 Μαρτίου 2020, και περιλαμβάνει 123 παρατηρήσεις. Τον Νοέμβριο του 2019, συγκεκριμένα 16-18 Νοεμβρίου 2019, το πολωνικό e-Service παρουσιάστηκε σε εκπαιδευτικούς της χώρας με σκοπό την σταδιακή συστηματική εφαρμογή του. Η περίοδος αυτή, λοιπόν, αποτελεί περίοδο γνωριμίας με την πλατφόρμα και εκπαίδευσης των χρηστών. Στη συνέχεια παρουσιάζονται περιγραφικά οι μετρικές σχετικές με Visitor Type Metrics και Activity Metrics με τα αντίστοιχα διαγράμματα κατανομών και θηκογραμμάτων.

	Unique visitors	Returning Visits	Visits	New Visits
count	123	123	123	123
mean	9.03	9.01	13.09	4.81
median	6.00	6.00	9.00	3.00
std	8.65	8.96	13.65	5.53
min	1.00	1.00	1.00	1.00
25%	4.00	3.00	4.00	2.00
75%	11.00	11.00	17.00	5.00
max	58.00	49.00	89.00	40.00

Πίνακας 5.1: Μέτρα διαφορετικών τύπων επισκεπτών

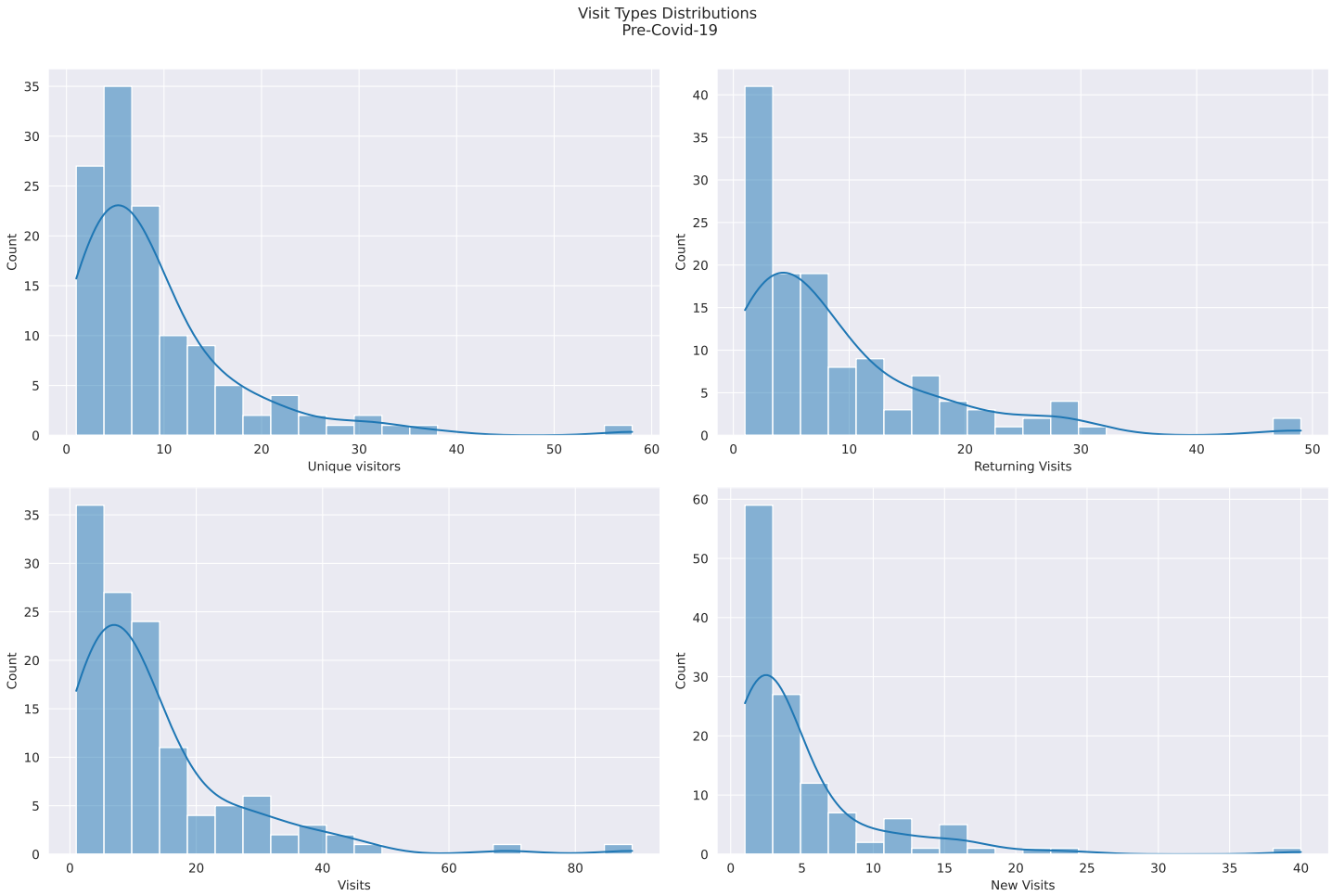
Από τον Πίνακα 5.1, τα διαγράμματα κατανομών και τα θηκογράμματα των Visitor Type Metrics (σελ. 53) προκύπτουν τα εξής:

- Παρατηρείται μη συστηματική χρήση της πλατφόρμας με χαμηλές τιμές σε όλες τις μεταβλητές και με μεμονωμένες μόνο ημέρες να εμφανίζουν μικρή, μη σημαντική, αύξηση. Αυτό φαίνεται και από τα θηκογραφήμματα τα οποία καταγράφουν τις ακραίες αυτές τιμές. Οι ημέρες αυτές ενδεχομένως να σχετίζονται με ημέρες εκπαίδευσης των χρηστών ή άλλου γεγονότος που δεν είναι γνωστό.

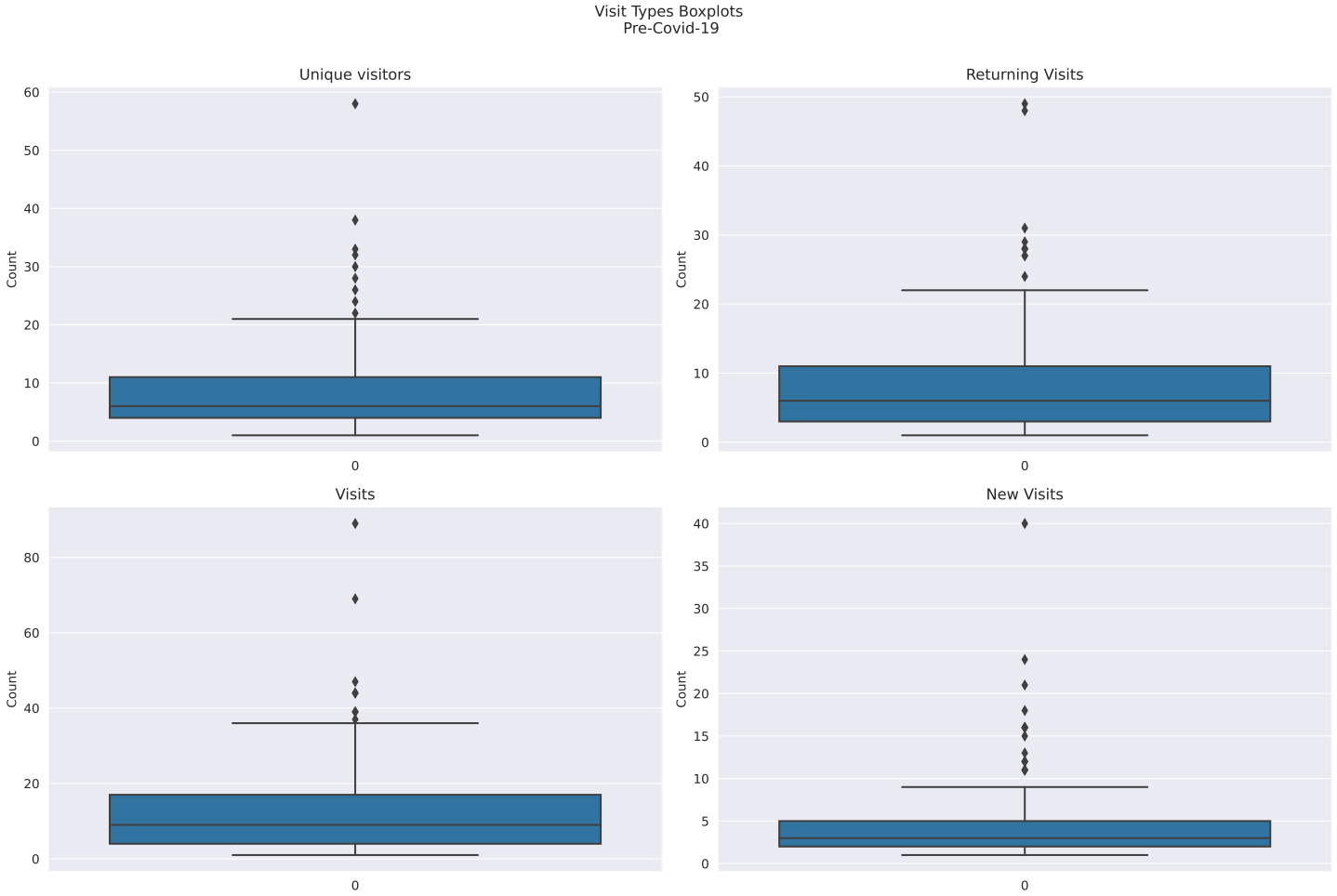
- Η μεταβλητή Visits εμφανίζει τον μεγαλύτερο μέσο όρο 13.09, ενώ οι μεταβλητές Returning Visits και Unique Visitors παρουσιάζουν ομοιότητες. Αντίθετα η New Visits έχει τον μικρότερο μέσο όρο (4.81) και μικρές καταγραφές. Συνεπάγεται πως δεν υπήρξε εισροή χρηστών και πρόκειται για τους ίδιους επισκέπτες να πραγματοποιούν συνδέσεις-επισκέψεις.

- Τα διαγράμματα κατανομών εμφανίζουν όλα θετική ασυμμετρία, που εξηγεί πως τις περισσότερες ημέρες καταγράφηκε πολύ μικρός αριθμός επισκεπτών για κάθε τύπο. Παράλληλα προκύπτει μεγάλη διασπορά που όμως επηρεάζεται από τις ακραίες τιμές. Συγκεκριμένα, για τα Visits καταγράφηκαν περισσότερες ημέρες με 20 ή λιγότερες επισκέψεις, εξαίρεση αποτελεί η 10η Δεκεμβρίου που συγκέντρωσε 89 επισκέψεις, για Returning Visits τις περισσότερες ημέρες λιγότερο από 10 συνδέσεις, για New Visits παρατηρήθηκαν τις περισσότερες ημέρες 0-5 νέες επισκέψεις κ.ο.κ..

Σημείωση: Στο διάστημα είναι το μοναδικό στο οποίο σημειώθηκαν και όλες οι ελλιπείς τιμές, που εν συνεχεία υπολογίσθηκαν με τη μέθοδο kNN Imputation. Για τη μεταβλητή New Visits συγκεκριμένα υπήρξαν 22/123 τιμές που χρειάστηκε να υπολογιστούν. Η μικρή αυτή μεροληψία είναι καλό να καταγράφεται για λόγους ερμηνείας των δεδομένων.



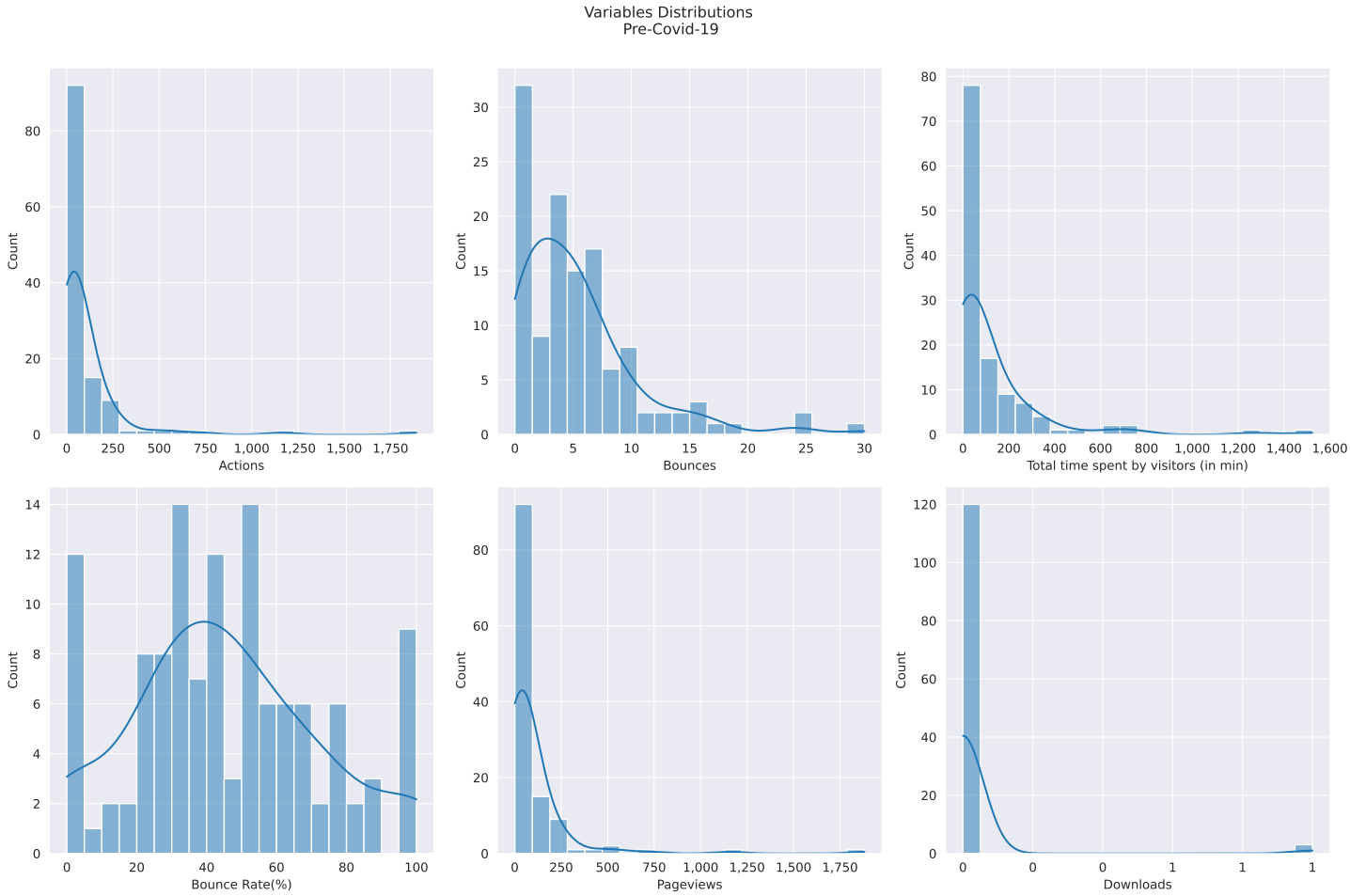
(α′) Κατανομές Visitor Type Metrics Pre Covid-19



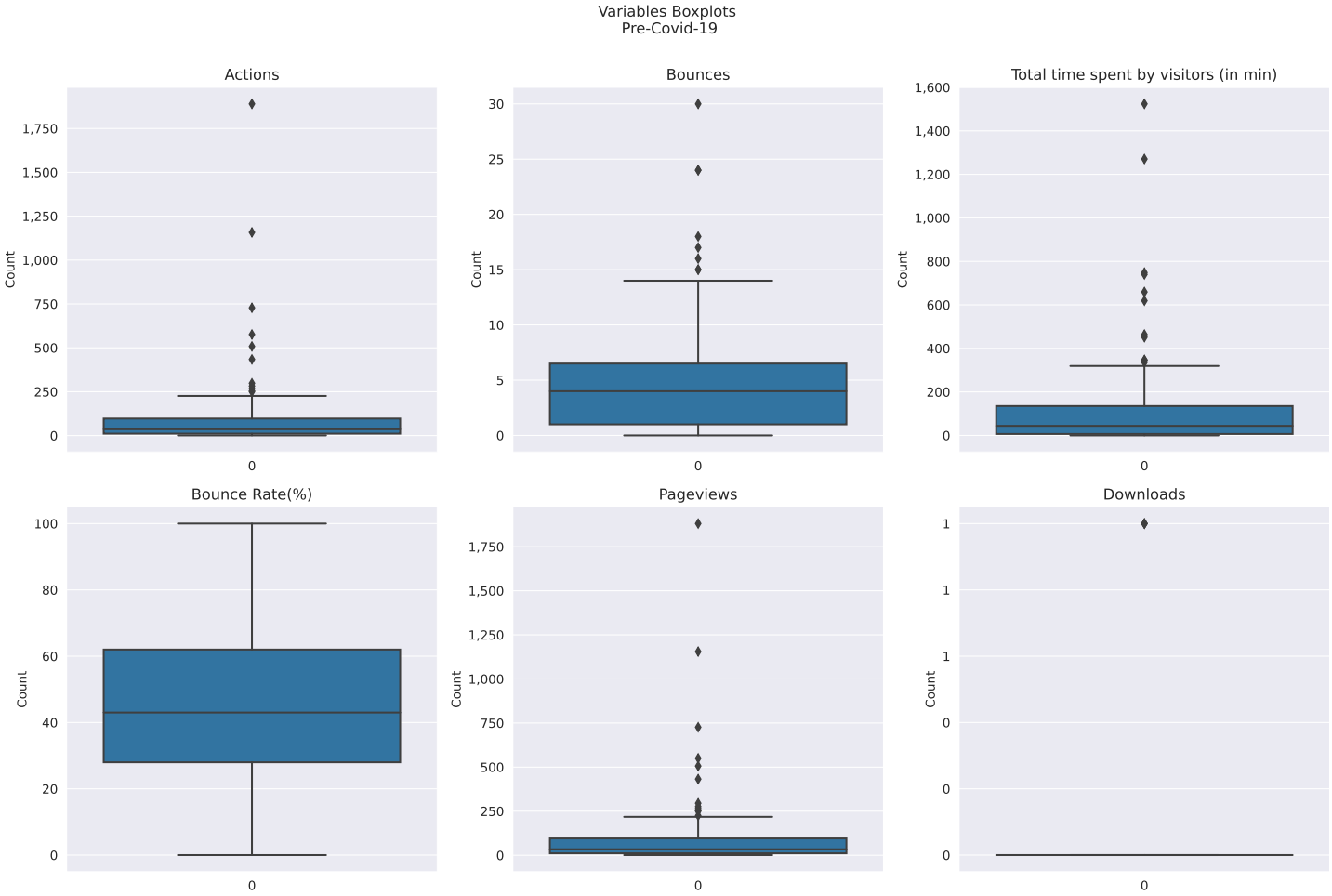
(β′) Θηκογράμματα Visitor Type Metrics Pre Covid-19

	Actions	Bounces	Total time spent by visitors (in min)	Bounce Rate(%)	Pageviews	Downloads
count	123	123	123	123	123	123
mean	101.16	5.29	121.91	44.91	100.31	0.02
median	36.00	4.00	43.97	43.00	34.00	0.00
std	221.14	5.20	222.82	26.79	219.78	0.15
min	1.00	0.00	0.00	0.00	1.00	0.00
25%	11.00	1.00	6.60	28.00	11.00	0.00
75%	97.00	6.50	135.04	62.00	95.50	0.00
max	1,890.00	30.00	1,523.98	100.00	1,881.00	1.00

Πίνακας 5.2: Περιγραφικός πίνακας μετρικών Activity Metrics Pre Covid-19



Σχήμα 5.5: Κατανομές Activity Metrics Pre Covid-19



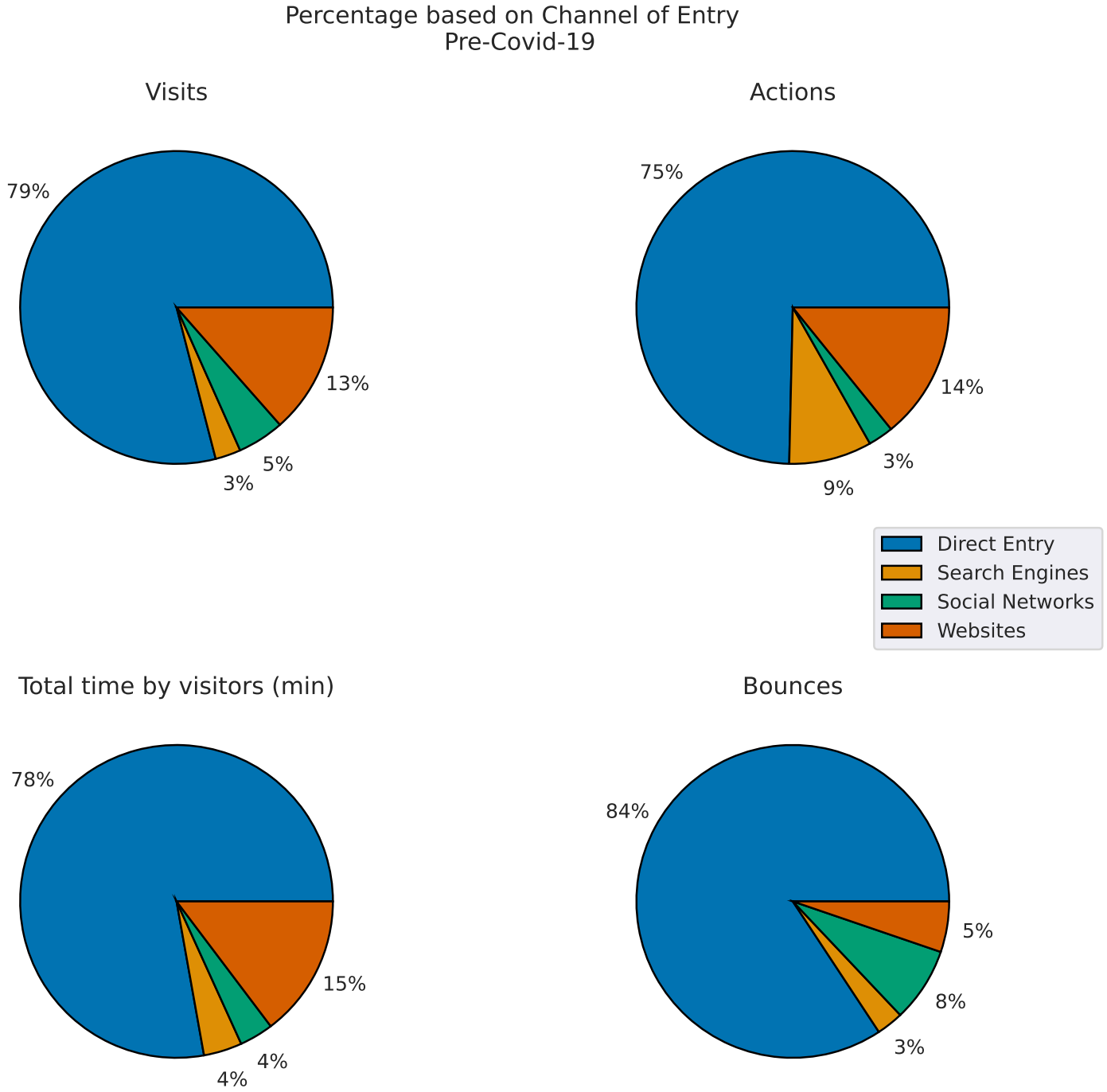
Σχήμα 5.6: Θηκογράμματα Activity Metrics Pre Covid-19

Αντίστοιχα, για τις μεταβλητές Activity Metrics προκύπτουν:

- Οι κατανομές εμφανίζουν θετική ασυμμετρία (εξαιρείται η Bounce Rate(%)) καθώς οι περισσότερες μεταβλητές καταγράφουν μικρές τιμές τις περισσότερες ημέρες. Φαίνεται να υπάρχει σημαντική διασπορά κυρίως για Actions, Total Time spent by visitors (in min), Pageviews αφού επηρεάζονται από ακραίες τιμές. Μικρότερη διασπορά εμφανίζουν Downloads και Bounces. Οι ακραίες τιμές που επηρεάζουν τις μεταβλητές είναι εμφανείς και στα θηκογράμματα, όπως συμβαίνει για την Actions, Total Time spent by visitors (in min).

- Αξιοσημείωτο είναι πως τα Downloads εμφανίζουν σχεδόν μηδενικές τιμές που σημαίνει πως η πλατφόρμα δεν χρησιμοποιήθηκε για λήψη αρχείου, ή δεν καταγράφηκε ή όλο το σχετικό περιεχόμενο ήταν διαθέσιμο στη σελίδα, επομένως δεν απαιτούνταν λήψη.

- Σχετικά με του Bounce Rate(%), δεν ακολουθεί γνωστή κατανομή κι εμφανίζει μεγαλύτερο εύρος τιμών με τον μέσο όρο (44.91%) και τη διάμεσο(43%) να βρίσκονται σχετικά κοντά αλλά επηρεάζεται από μεγάλες διακυμάνσεις.



Σχήμα 5.7: Κανάλια εισόδου Pre Covid-19

Από το Σχήμα 5.7, φαίνεται πως οι περισσότερες επισκέψεις προέρχονται από απευθείας σύνδεση (79%), δηλαδή είτε είχαν αποθηκευμένο το σύνδεσμο εισόδου, είτε τον είχαν απομνημονεύσει, κι ακολουθεί καταγραφή επισκέψεων από άλλες ιστοσελίδες (13%). Οι υπόλοιπες μετρικές είναι ακολουθούν ανάλογη συμπεριφορά όπως των επισκέψεων ανά κανάλι εισόδου.

5.1.2 Περίοδος Wave 1

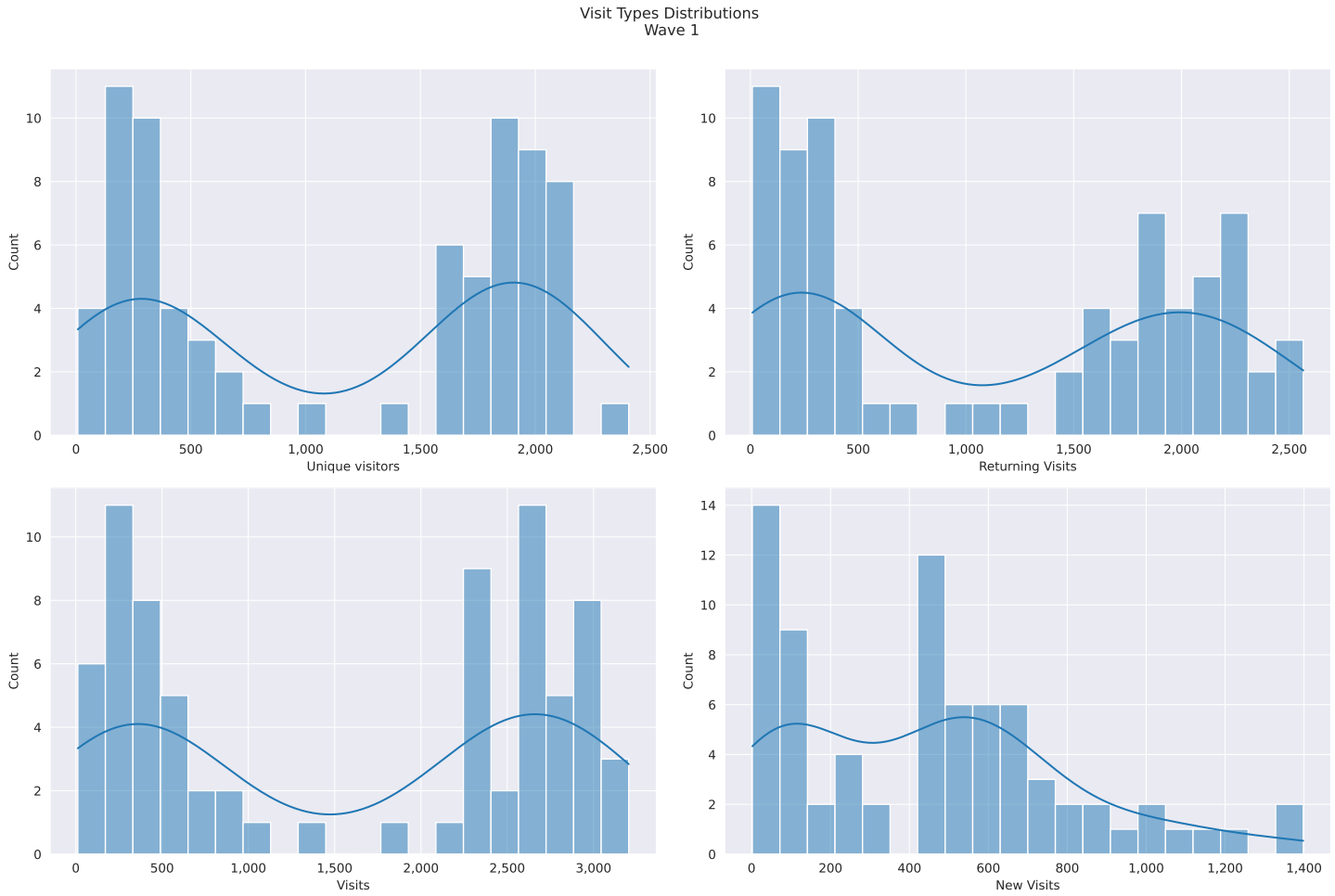
Η περίοδος Wave 1 αναφέρεται στις ημερομηνίες 10 Μαρτίου 2020 - 24 Μαΐου 2020, και περιλαμβάνει 76 παρατηρήσεις. Οι ημερομηνίες συμπίπτουν με την έναρξη και λήξη του υποχρεωτικού εγκλεισμού λόγω Covid-19 και της υποχρεωτικής τηλεκπαίδευσης στην Πολωνία. Την περίοδο αυτή το Up2U αποτέλεσε βασική πλατφόρμα για την εξυπηρέτηση των απομακρυσμένων εκπαιδευτικών αναγκών. Ακολουθεί ανάλυση για τις μετρικές σε αυτή την περίοδο.

	Unique visitors	Returning Visits	Visits	New Visits
count	76.00	76.00	76.00	76.00
mean	1,150.72	1,135.05	1,587.28	452.22
median	1,595.00	1,189.50	2,217.50	464.00
std	822.78	903.03	1,165.72	351.10
min	8.00	9.00	12.00	2.00
25%	278.50	255.75	360.25	98.00
75%	1,917.50	1,968.75	2,707.75	649.00
max	2,407.00	2,565.00	3,204.00	1,398.00

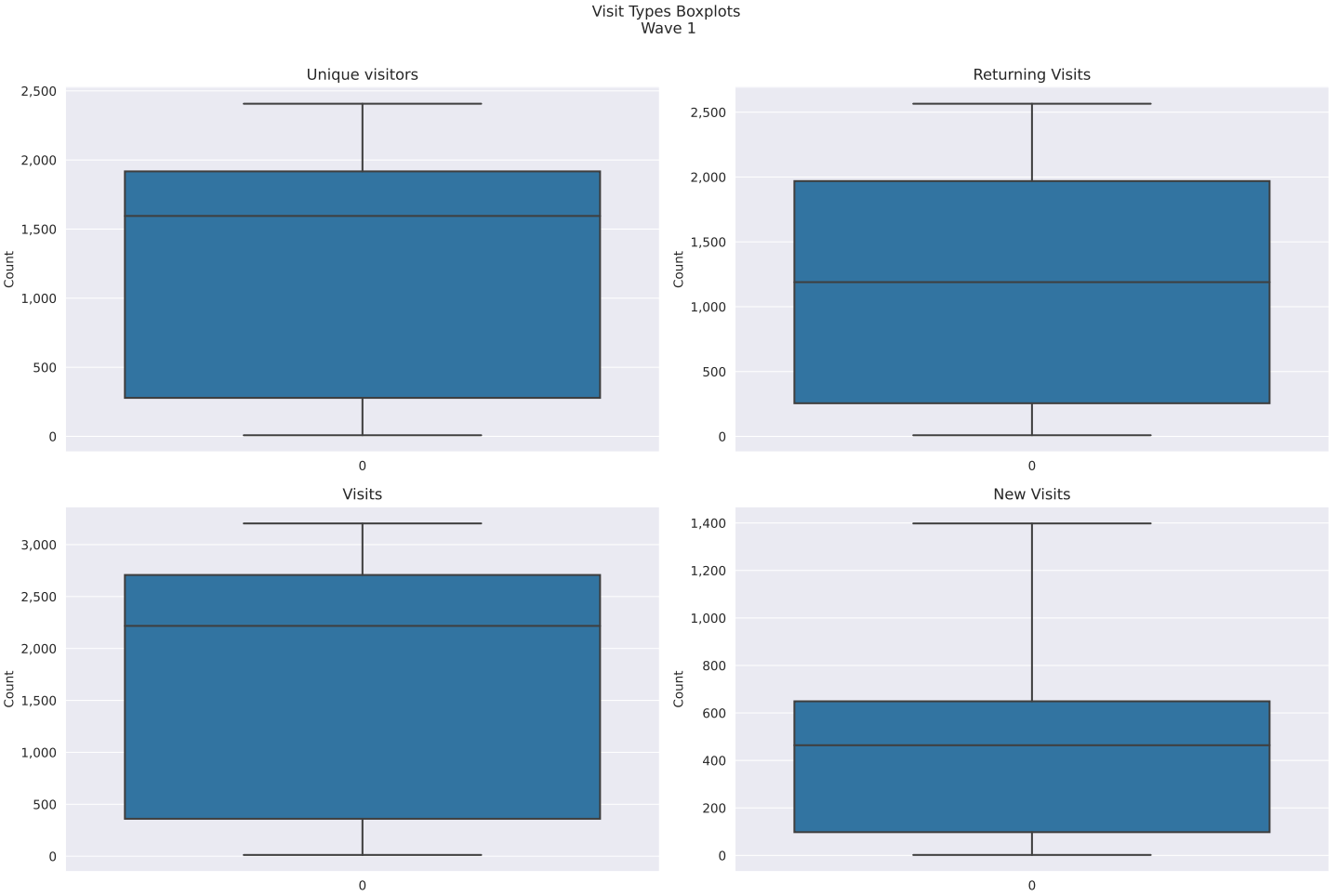
Πίνακας 5.3: Μέτρα διαφορετικών τύπων επισκεπτών Wave 1

Από την Πίνακα 5.3, τα διαγράμματα κατανομών και θηκογραμμάτων της ακόλουθης σελίδας 58, συμπεραίνεται ότι:

- Όπως έδειξε και το Σχήμα 5.1 ημερήσιων επισκέψεων, το e-Service χρησιμοποιείται με συστηματικό τρόπο καθημερινά στο διάστημα αυτό. Επομένως δεν εμφανίζονται ακραίες τιμές, ωστόσο υπάρχει αξιοσημείωτη διασπορά και εύρος τιμών.
- Οι καμπύλες των Visits, Returning Visits, Unique Visitors εμφανίζουν διτροπική (bimodal) κατανομή, δηλαδή μέρες χαμηλής και υψηλής επισκεψιμότητας. Λόγω της καθημερινής χρήσης η κατανομή αυτή ενδεχομένως να δημιουργείται λόγω των διαφορετικών ημερών της εβδομάδας με περισσότερη ή λιγότερη χρήση (π.χ. καθημερινές έναντι σαββατοκύριακου) ή άλλων γεγονότων (π.χ. ημέρες εξέτασης).
- Η New Visits εμφανίζει διτροπική κατανομή με θετική ασυμμετρία καθώς συγκέντρωσε αυξήθηκαν σταδιακά οι νέοι επισκέπτες μέχρι 25 Μαρτίου που σημείωσε την υψηλότερη τιμή της 1,398. Αυτή είναι και η υψηλότερη τιμή της μεταβλητής σε όλο το σύνολο δεδομένων. Συνεπάγεται πως το e-Service την ημέρα εκείνη προσέλκυσε τον μεγαλύτερο δυνατό αριθμό νέων χρηστών που ήταν στη δυνατότητά του.
- Παρατηρείται σημαντική αύξηση αύξηση σε όλες τις μεταβλητές (παρά τον μικρότερο αριθμό παρατηρήσεων) που δείχνει καθιέρωση της πλατφόρμας ως βασικό μέσο τηλεκπαίδευσης.



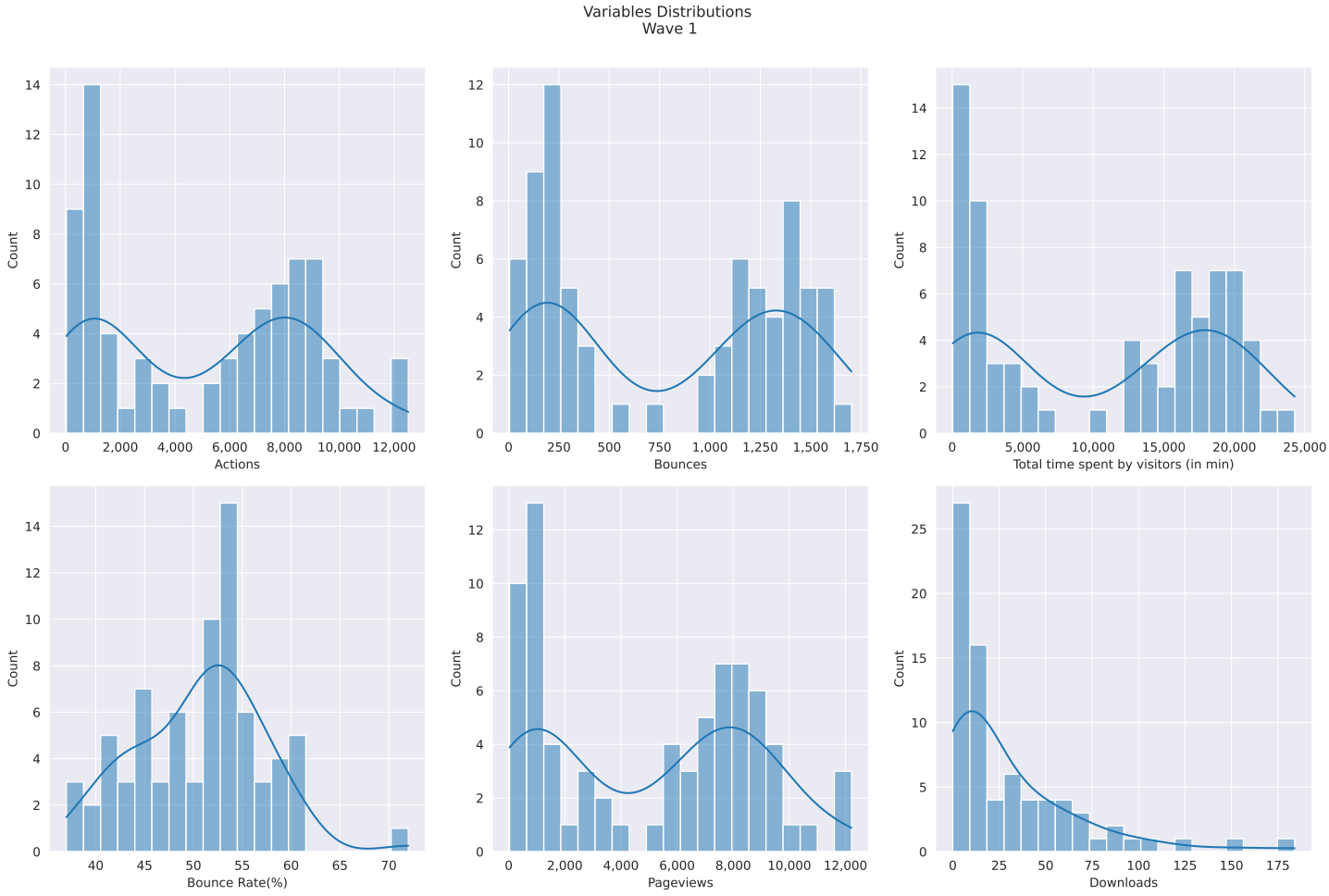
(α′) Κατανομές Visitor Type Metrics Wave 1



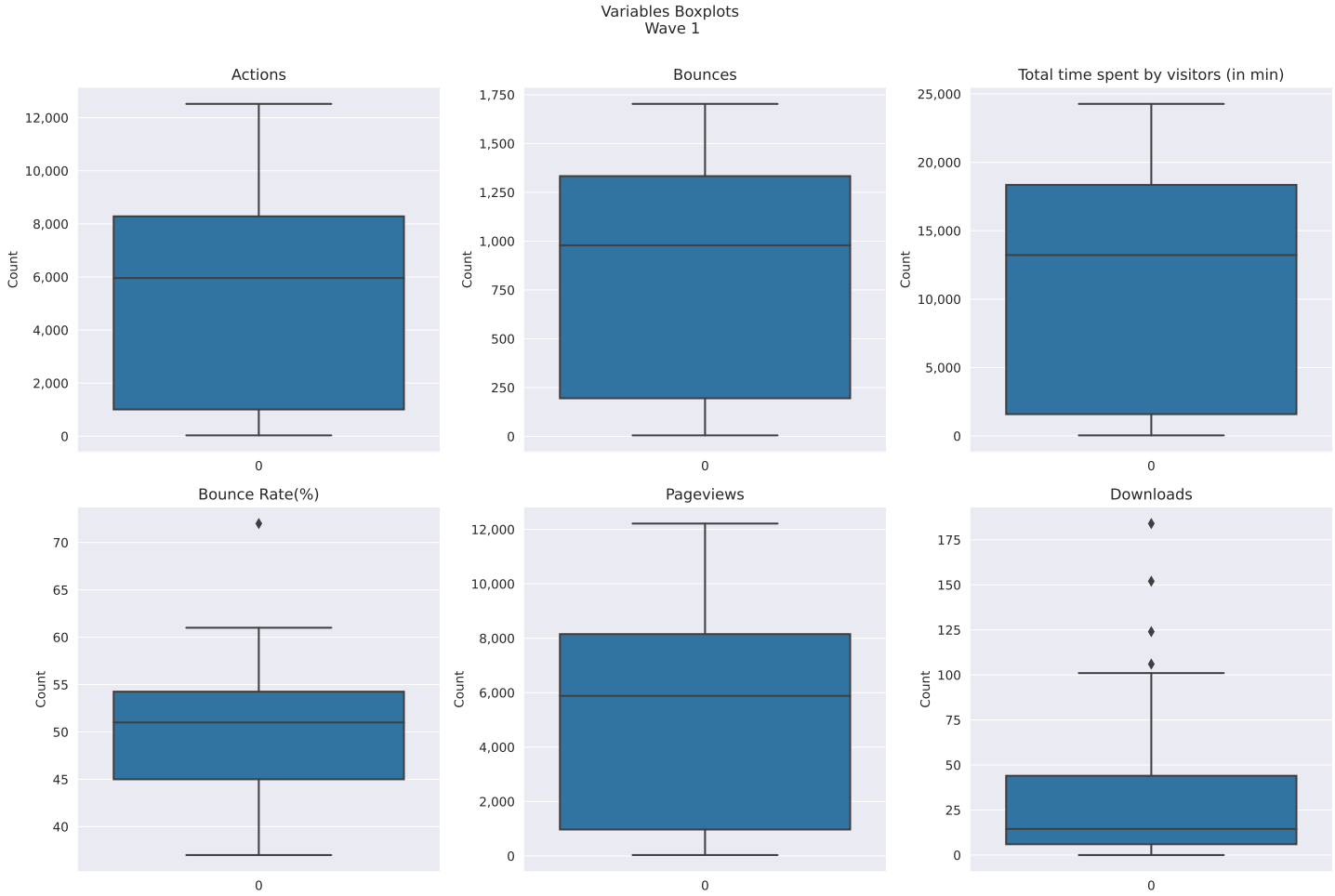
(β′) Θηκογράμματα Visitor Type Metrics Wave 1

	Actions	Bounces	Total time spent by visitors (in min)	Bounce Rate(%)	Pageviews	Downloads
count	76	76	76	76	76	76
mean	5,113.80	787.75	10,648.44	50.63	5,024.09	30.20
median	5,958.50	978.50	13,221.78	51.00	5,880.50	14.50
std	3,790.95	587.64	8,215.13	6.57	3,721.64	36.73
min	32.00	5.00	41.35	37.00	32.00	0.00
25%	1,009.25	195.00	1,593.03	45.00	970.75	6.00
75%	8,282.75	1,333.00	18,352.31	54.25	8,151.00	44.00
max	12,520.00	1,703.00	24,272.68	72.00	12,214.00	184.00

Πίνακας 5.4: Περιγραφικός πίνακας μετρικών Activity Metrics



Σχήμα 5.9: Κατανομές Activity Metrics Wave 1

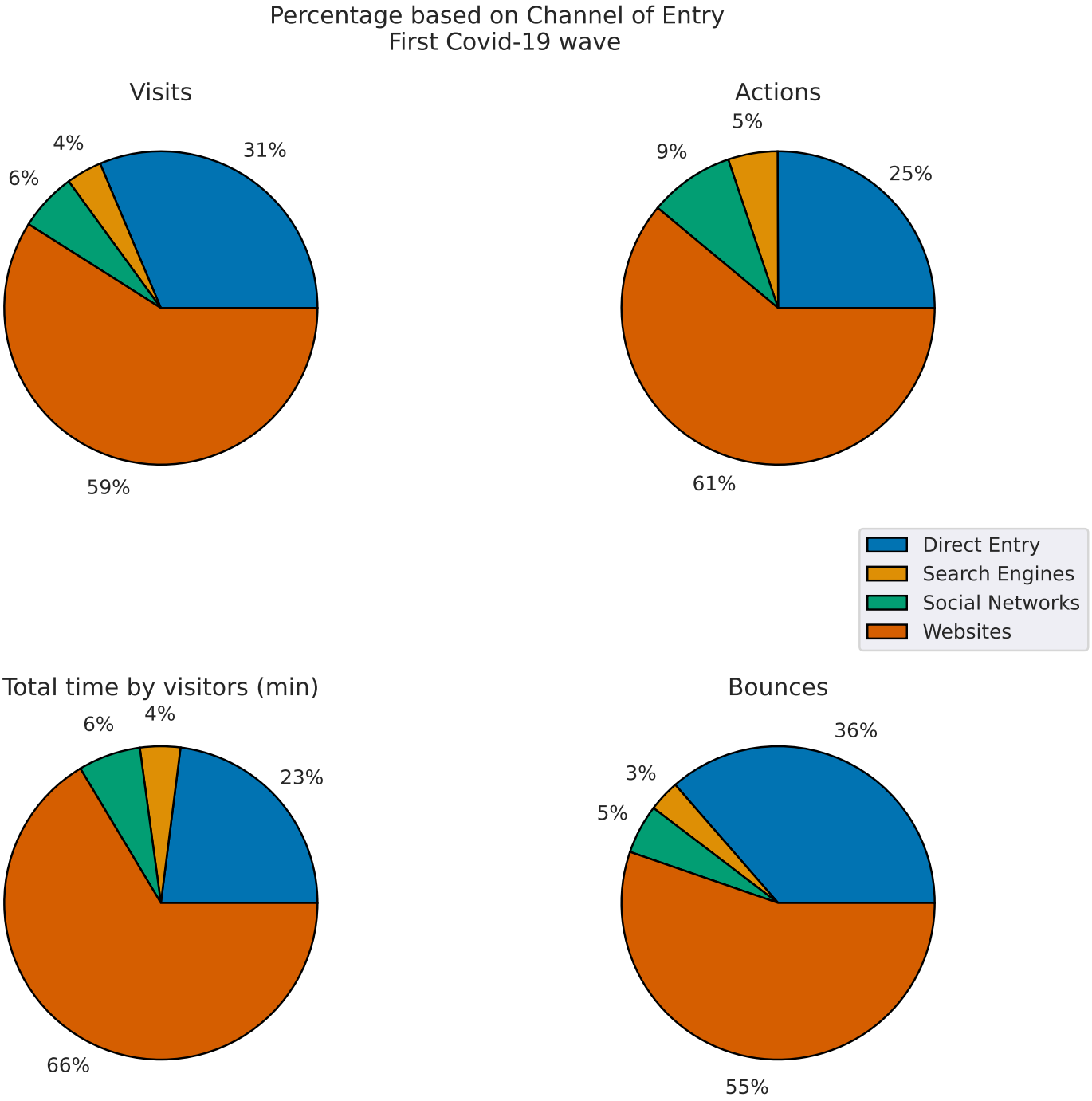


Σχήμα 5.10: Θηκογράμματα Activity Metrics Wave 1

Ομοίως για τα Activity Metrics της περιόδου Wave 1:

- Οι μετρικές Actions, Bounces, Pageviews, Total Time spent by visitors (in min) ακολουθούν διτροπική κατανομή, ακολουθώντας αυτές των Visitor Type Metrics, με μεγάλη διασπορά. Αντίθετα η Downloads παρουσιάζει θετική ασυμμετρία, μεγάλη διασπορά, και η Bounce Rate(%) που επηρεάζεται από ακραίες τιμές που συγκεντρώνει το πλήθος των τιμών της στο κέντρο με μικρή διασπορά, ωστόσο δεν ακολουθεί γνωστή κατανομή. Δεν προκύπτουν ακραίες τιμές με εξαίρεση την Bounce Rate(%) και την Downloads.

- Η Downloads, και πάλι, εμφανίζει μικρές τιμές ($\max = 184, \text{mean} = 30.20$), ενισχύοντας την υπόθεση πως το e-Service δεν χρησιμοποιήθηκε σημαντικά για λήψη αρχείων. Ακόμη, ο συνολικός χρόνος Total Time spent by visitors (in min) αντιστοιχεί σε περίπου 6.71 λεπτά ανά επίσκεψη κατά μέσο όρο. Τέλος, η σχεδόν ταύτιση τιμών των Actions και Pageviews δηλώνει πως πραγματοποιούνταν μία ενέργεια ανά σελίδα περίπου.



Σχήμα 5.11: Κανάλια εισόδου Wave 1

Στην περίοδο Wave 1 το κυρίαρχο μέσο εισόδου αποτελούν οι ιστοσελίδες, με την άμεση σύνδεση να πέφτει στο 31%. Αυτό δείχνει μια σημαντική αλλαγή στην συμπεριφορά των επισκεπτών και στη χρήση της πλατφόρμας που ίσως να εξηγείται στο ότι χρησιμοποιήθηκαν παραπομπές από σελίδες των εκπαιδευτικών ή του σχολείου. Προφανώς, υπάρχει αναλογία μεταξύ επισκέψεων ανά κανάλι εισόδου με Actions, Total Time spent by visitors (in min) και Bounces.

5.1.3 Περίοδος Wave 2

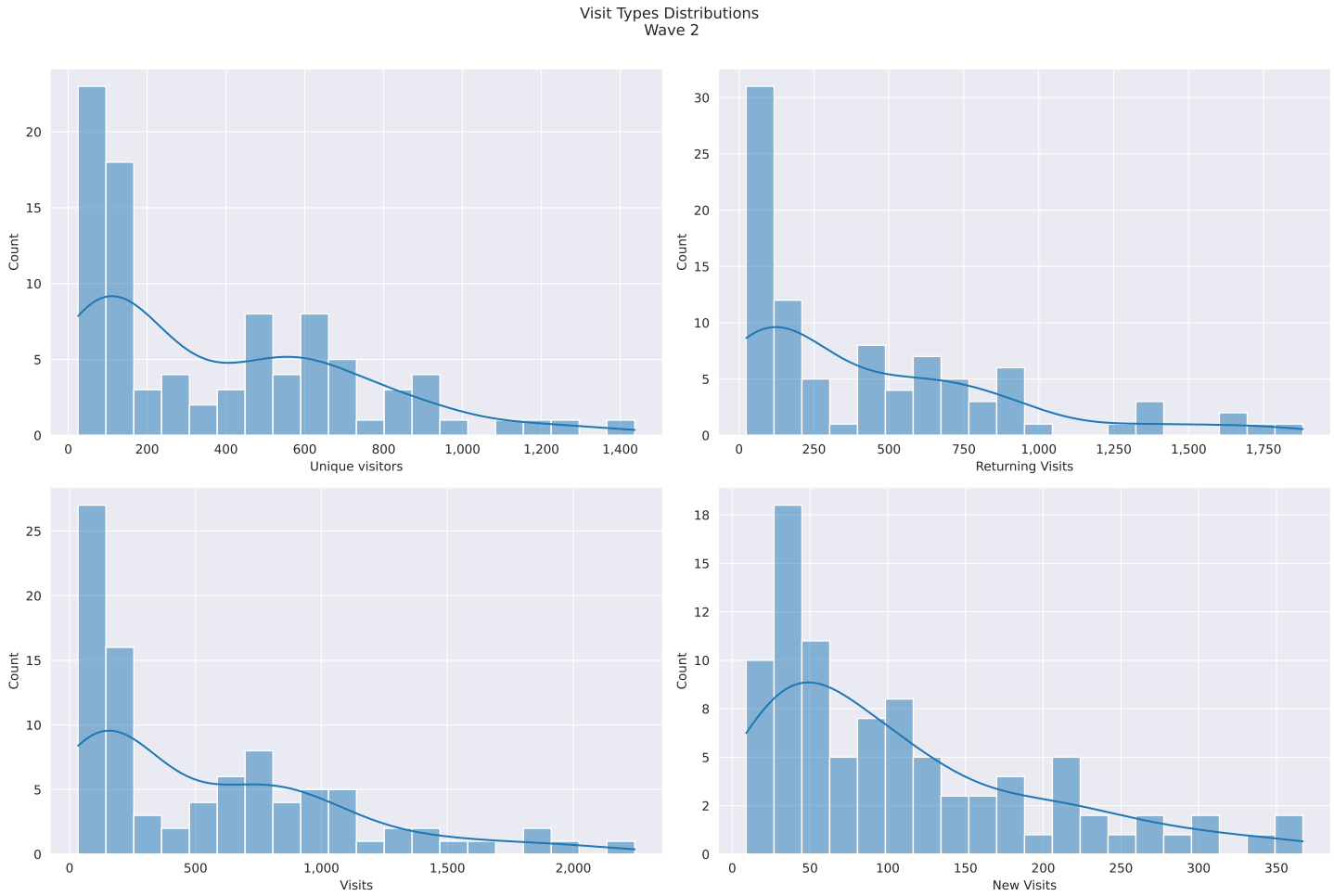
Η περίοδος Wave 2 αναφέρεται στο διάστημα 19 Οκτωβρίου 2020 - 17 Ιανουαρίου 2021, και περιλαμβάνει 91 παρατηρήσεις. Αποτελεί το δεύτερο πανδημικό κύμα Covid-19 που οδήγησε την πολωνική κυβέρνηση στο σταδιακό κλείσιμο των βαθμίδων εκπαίδευσης και την επαναφορά της τηλεκπαίδευσης. Σε σύγκριση με το πρώτο κύμα, αυτό έγινε σταδιακά: 19 Οκτωβρίου έγινε επανεκκίνηση τηλεκπαίδευσης για τη δευτεροβάθμια και τριτοβάθμια εκπαίδευση, 26 Οκτωβρίου για τις μεγαλύτερες τάξεις της πρωτοβάθμιας εκπαίδευσης και, τέλος, 9 Νοεμβρίου για τις τάξεις πρώτης έως τρίτης δημοτικού.

	Unique visitors	Returning Visits	Visits	New Visits
count	91.00	91.00	91.00	91.00
mean	381.30	443.03	551.71	108.68
median	255.00	256.00	344.00	81.00
std	336.96	452.62	520.66	89.96
min	25.00	24.00	34.00	9.00
25%	95.00	84.50	122.50	35.50
75%	602.50	674.00	836.50	159.50
max	1,437.00	1,881.00	2,244.00	367.00

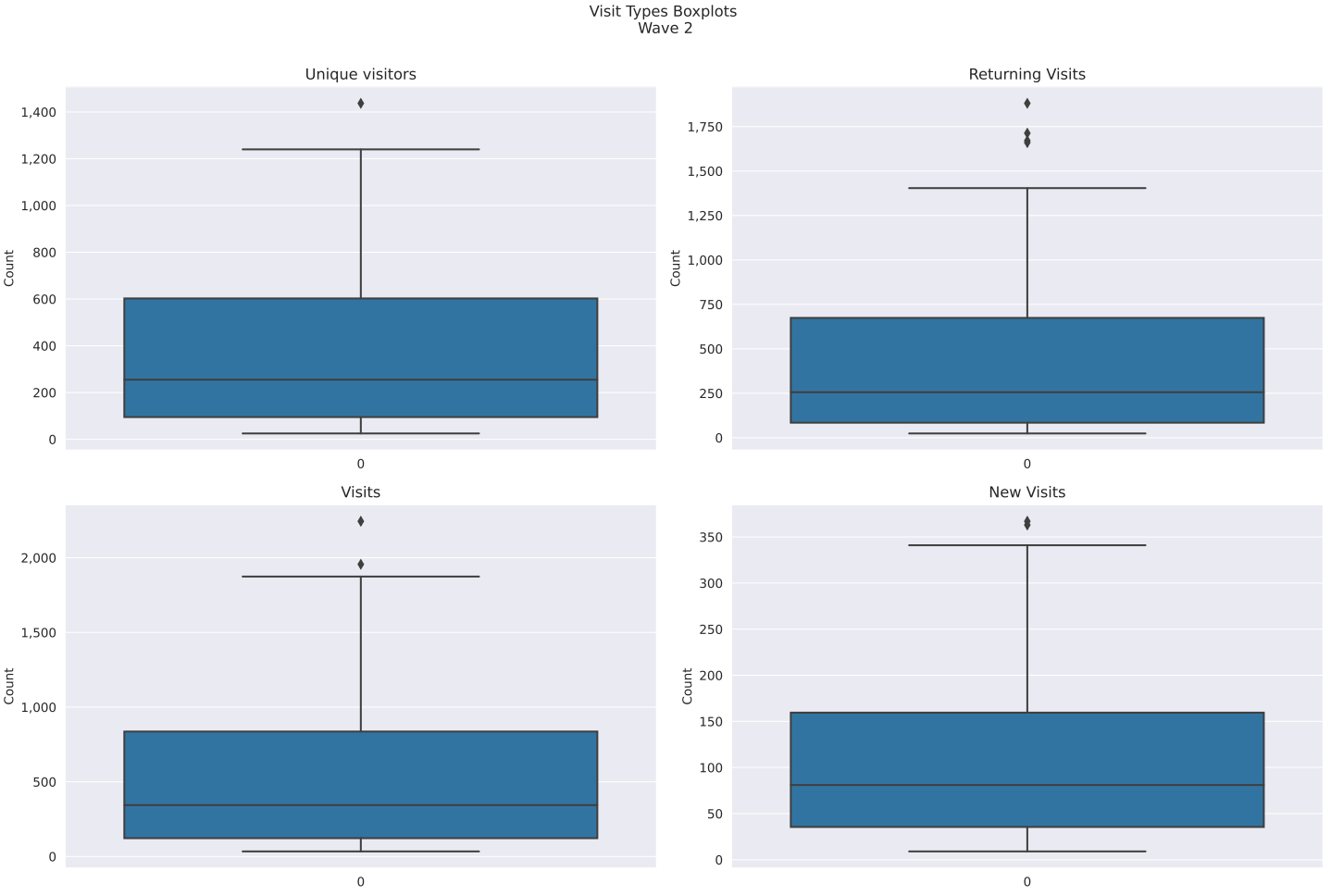
Πίνακας 5.5: Μέτρα διαφορετικών τύπων επισκεπτών Wave 2

Από τον παραπάνω πίνακα και τα διαγράμματα κατανομής και θηκογράμματα της σελίδας 63 προκύπτουν:

- μειωμένη χρήση σε σύγκριση με την περίοδο Wave 1 ωστόσο οι τιμές κυμαίνονται σε υψηλά επίπεδα. Σε όλες τις μεταβλητές υπάρχουν ακραίες τιμές που συμπίπτουν, μερικώς, με τις προαναφερθείσες ημερομηνίες. Η μεταβλητή New Visits παρουσιάζει θετική ασυμμετρία
- Οι καμπύλες των Unique Visitors, Returning Visits, Visits εμφανίζουν θετική ασυμμετρία, ωστόσο με μεγαλύτερη διαφορά μεταξύ των κορυφών σε σχέση με το Wave 1, εμφανώς επηρεασμένη από από τις ακραίες τιμές. Ακόμη, οι υψηλές τυπικές αποκλίσεις δείχνουν μεγάλη διασπορά.
- Παρουσιάζεται μια συστηματική, σχετικά, χρήση της πλατφόρμας που ωστόσο επηρεάζεται σημαντικά από ακραίες τιμές. Οι Returning Visits φαίνεται να είναι οι βασικοί επισκέπτες που σημαίνει πως δεν υπήρξε σημαντική εισροή νέων επισκεπτών σε σχέση με την περίοδο Wave 1, αν και υπάρχει παρουσία αυτών, κυρίως κοντά σε ημέρες που αφορούν την έναρξη τηλεκπαίδευσης σε κάποια από τις βαθμίδες.



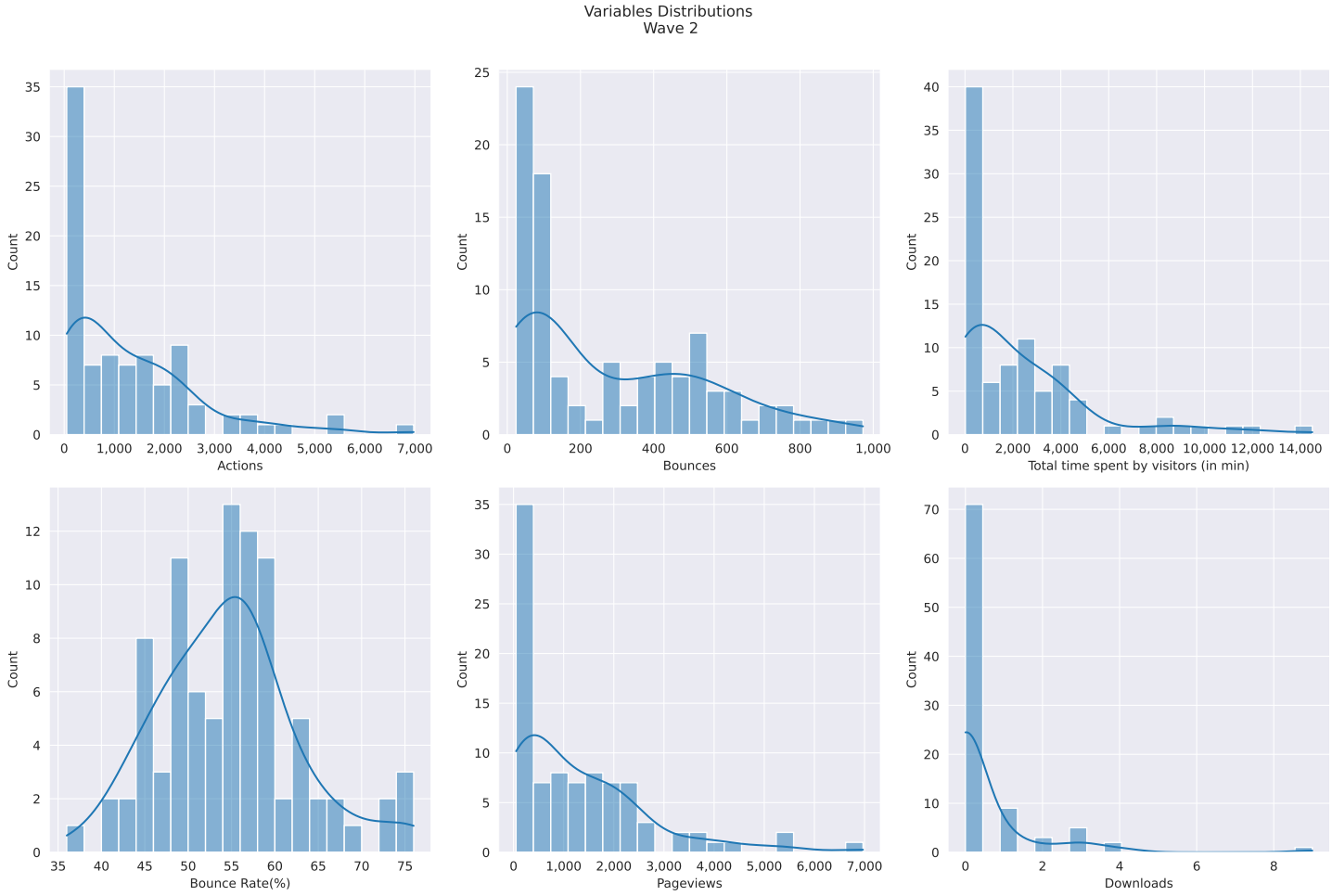
(α′) Κατανομές Visitor Type Metrics Wave 2



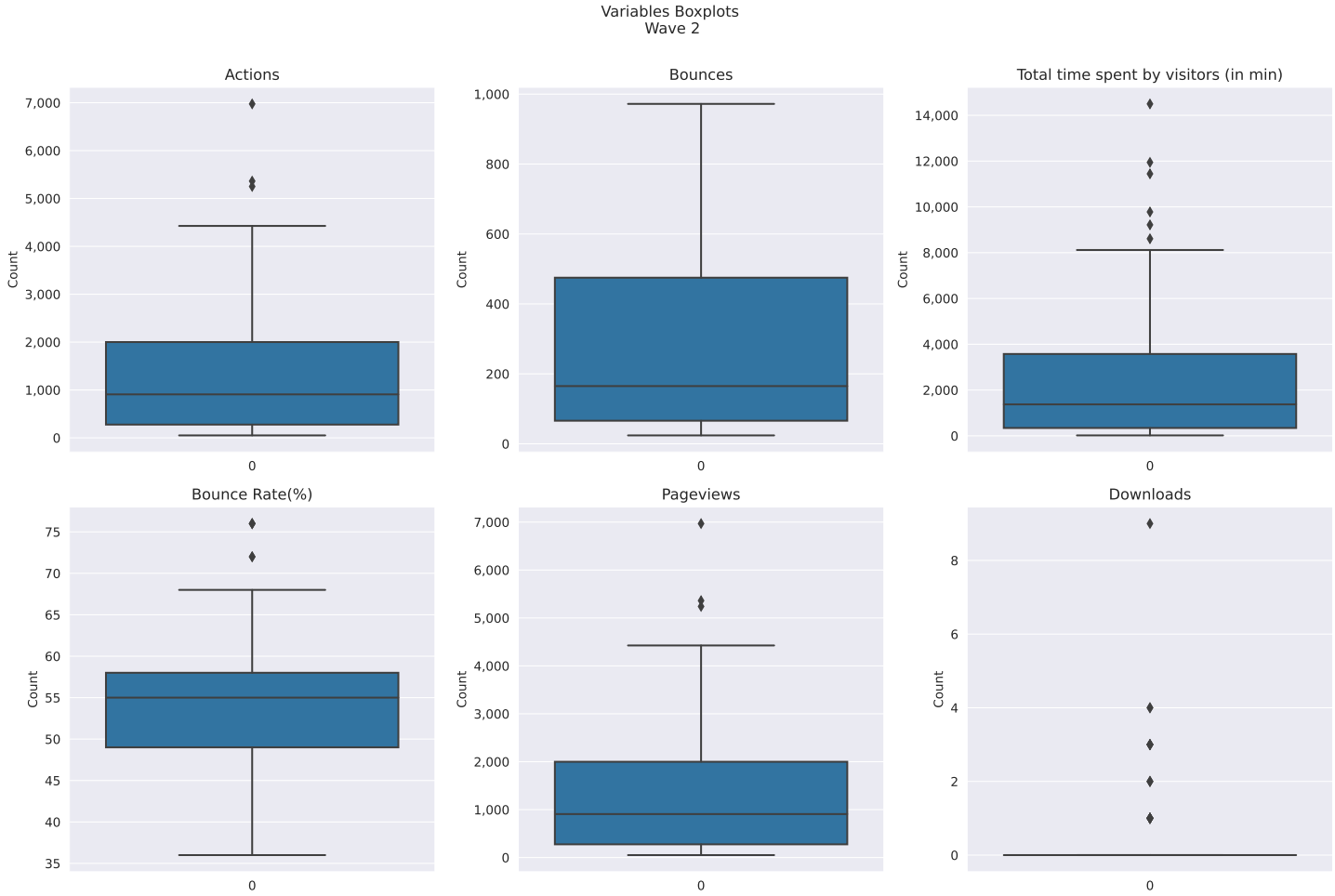
(β′) Θηκογράμματα Visitor Type Metrics Wave 2

	Actions	Bounces	Total time spent by visitors (in min)	Bounce Rate(%)	Pageviews	Downloads
count	91	91	91	91	91	91
mean	1,313.79	285.66	2,369.91	54.57	1,310.73	0.52
median	908.00	165.00	1,373.73	55.00	907.00	0.00
std	1,355.14	254.26	2,914.46	8.04	1,353.42	1.31
min	51.00	24.00	19.13	36.00	51.00	0.00
25%	277.00	66.00	343.24	49.00	276.00	0.00
75%	2,001.00	475.00	3,575.10	58.00	1,998.50	0.00
max	6,976.00	972.00	14,498.25	76.00	6,969.00	9.00

Πίνακας 5.6: Περιγραφικός πίνακας μετρικών Activity Metrics Wave 2



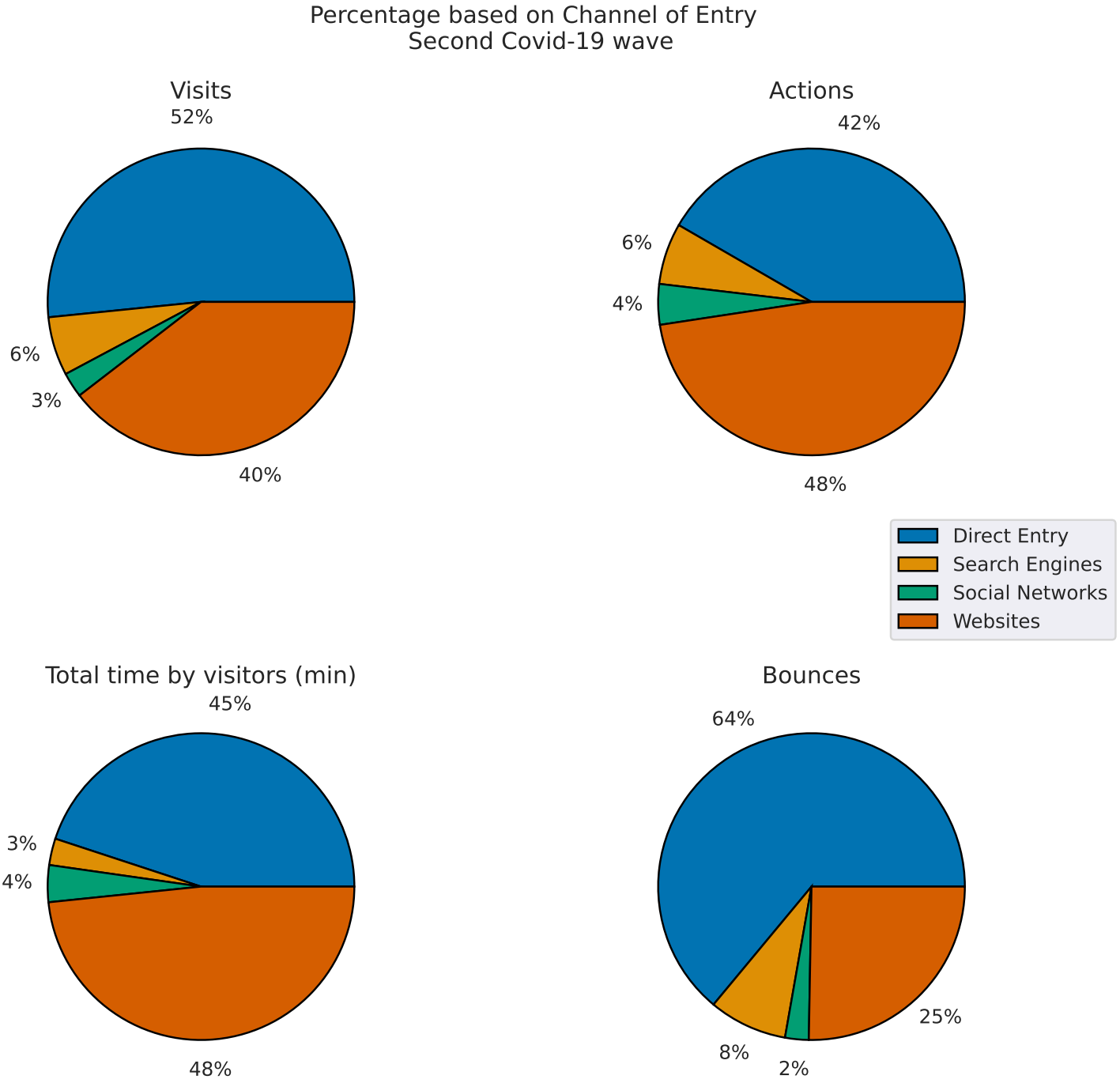
Σχήμα 5.13: Κατανομές Visitor Type Metrics Wave 2



Σχήμα 5.14: Θηκογράμματα Visitor Type Metrics Wave 2

Ομοίως για τις μετρικές Activity Metrics:

- Οι τιμές των μετρικών ακολουθούν την πτώση των επισκέψεων και εμφανίζουν μειωμένες τιμές σε σχέση με το Wave 1, σημαντική διασπορά, θετική ασυμμετρία, που επηρεάζεται από ακραίες τιμές. Εξαίρεση αποτελεί η Bounce Rate(%) που τείνει να γίνει κανονική.
- Οι τιμές των Downloads έπεσαν σε Pre Covid-19 τιμές, δείχνοντας πως δεν χρησιμοποιήθηκαν ιδιαίτερα οι λήψεις αρχείων ή πως το επιθυμητό περιεχόμενο βρισκόταν επισυναπτόμενο στη σελίδα. Από τις τιμές Total Time spent by visitors (in min) προκύπτει πως ο χρόνος παραμονής ήταν περίπου 4.3 λεπτά ανά επίσκεψη.
- Κι εδώ υπάρχει ταύτιση τιμών μεταξύ των μετρικών Actions και Pageviews, δηλαδή αναλογία 1:1, μία ενέργεια σε μία σελίδα.



Σχήμα 5.15: Κανάλια εισόδου Wave 2

Σχετικά με το κανάλια εισόδου για την περίοδο Wave 2, οι επισκέψεις από Direct Entry, επανήλθε στην πρώτη θέση ως κύριο μέσο σύνδεσης. Η είσοδος μέσω Websites αν και σημείωσε 19% πτώση, παρέμεινε σε υψηλά επίπεδα της τάξεως του 40%. Ενδιαφέρον έχει πως οι επισκέπτες από Websites, φαίνεται να έκαναν αναλογικά περισσότερα Actions και να περνούν περισσότερο χρόνο στην υπηρεσία. Για τα κανάλια εισόδου Social Networks, Search Engines δεν υπήρξε αξιοσημείωτη διαφορά.

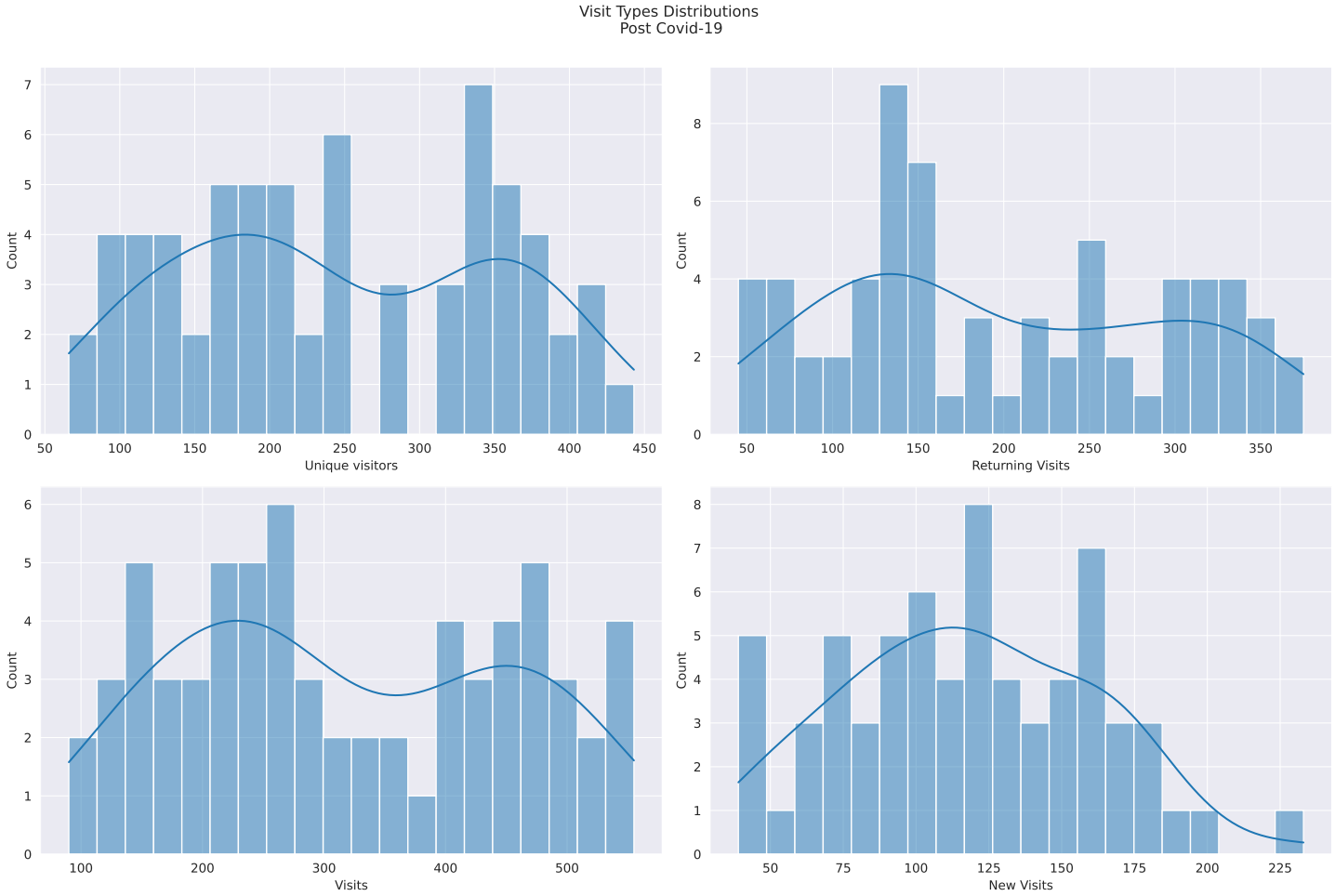
5.1.4 Περίοδος Post Covid-19

Η τελευταία περίοδος διερεύνησης Post Covid-19 αναφέρεται στις ημερομηνίες 18 Ιανουαρίου 2021 - 25 Μαρτίου 2021 και περιλαμβάνει 67 παρατηρήσεις. Η ονομασία δεν παραπέμπει σε λήξη της πανδημίας αλλά σε λήξη της υποχρεωτικής τηλεκπαίδευσης λόγω Covid-19 και επαναλειτουργία των δια ζώσης μαθημάτων. Η ημερομηνία 25 Μαρτίου 2021 αποτελεί και τελευταία ημερομηνία που το σύστημα του πολωνικού e-Service κατέγραψε οποιαδήποτε παρατήρηση λόγω λήξης του προγράμματος.

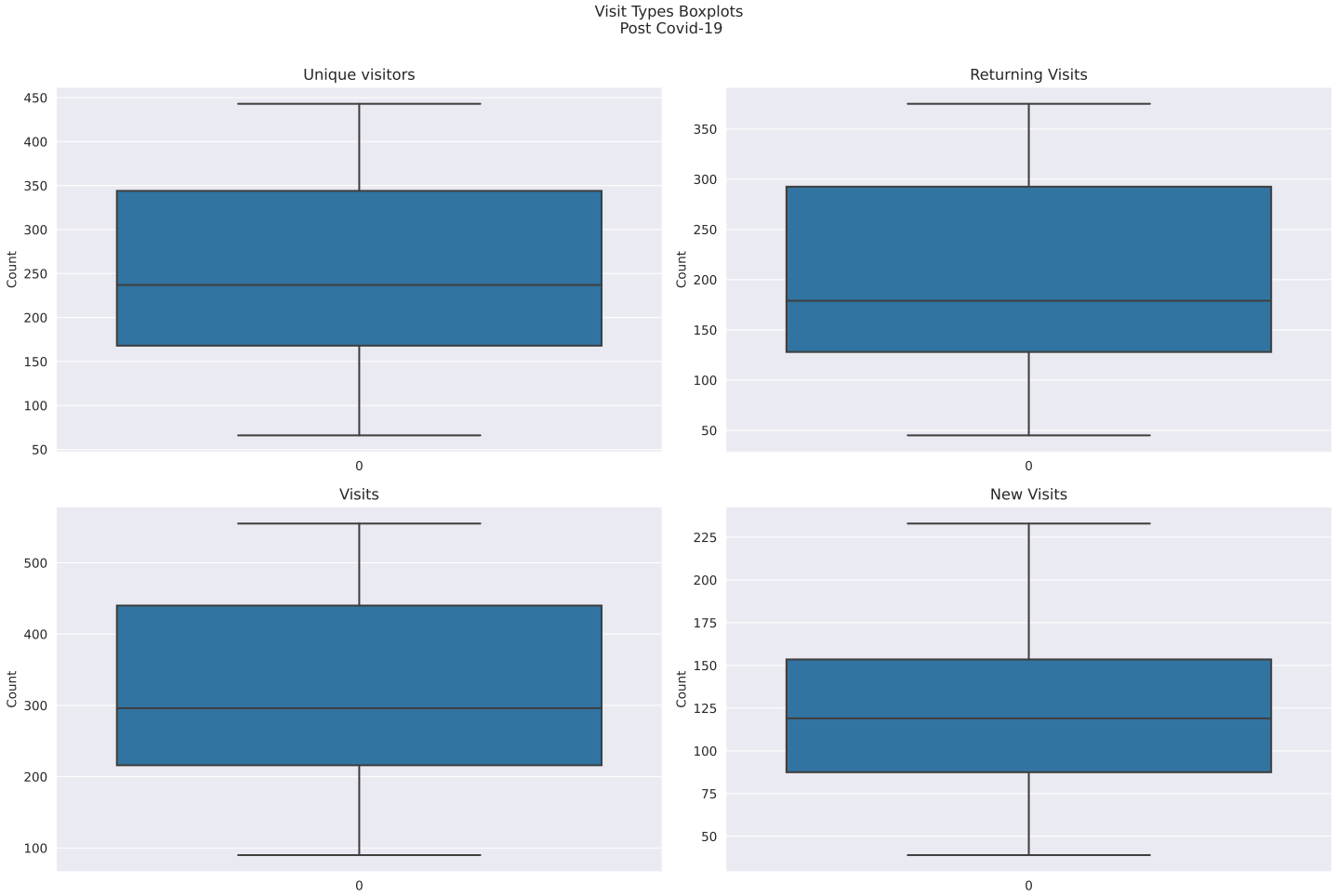
	Unique visitors	Returning Visits	Visits	New Visits
count	67	67	67	67
mean	246.81	199.06	317.18	118.12
median	237.00	179.00	296.00	119.00
std	105.73	96.88	134.88	43.14
min	66.00	45.00	90.00	39.00
25%	168.00	128.00	216.00	87.50
75%	344.00	292.50	440.00	153.50
max	443.00	375.00	555.00	233.00

Πίνακας 5.7: Μέτρα διαφορετικών τύπων επισκεπτών Post Covid-19

- Η περίοδος φαίνεται να χαρακτηρίζεται από μειωμένες τιμές σε σύγκριση με τα δύο κύματα πανδημίας, όμως εμφανίζει περιοδικότητα ως προς την επισκεψιμότητα, όπως δείχνει το Σχήμα 5.1, υψηλότερη και συστηματικότερη σε σχέση με την περίοδο Pre Covid-19, χωρίς ύπαρξη ακραίων τιμών.
- Το προηγούμενο φαίνεται να επηρεάζει και τις καμπύλες κατανομών που ακολουθούν πολυτροπική κατανομή (multimodal) και έχουν μεγάλες διασπορές. Συγκεκριμένα οι Unique Visitors, Visits τείνουν να γίνουν διτροπικές, ενώ οι New Visits και Returning Visits δεν ακολουθούν κάποια γνωστή κατανομή, ωστόσο υπάρχει σταθερή είσοδος νέων επισκεπτών.



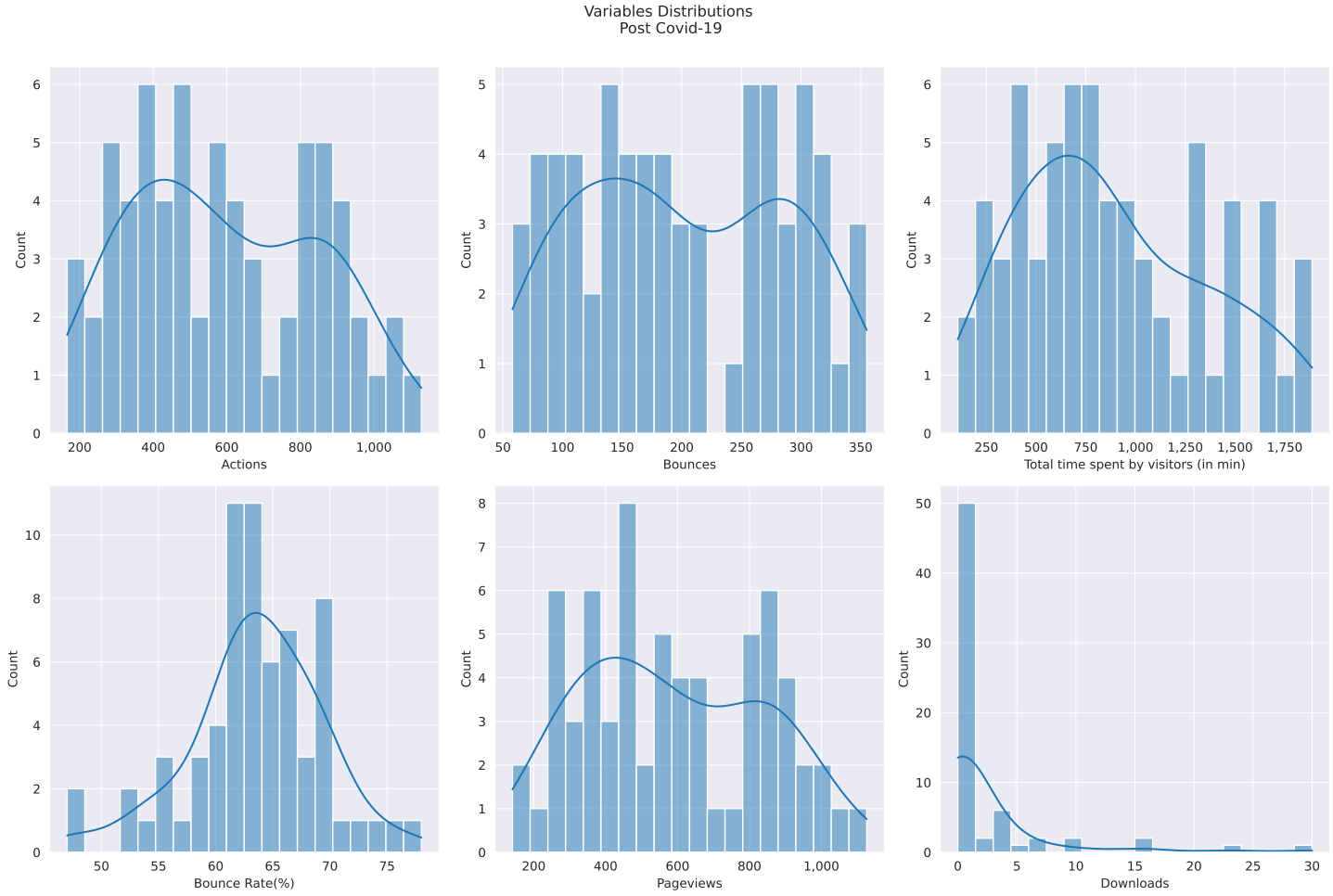
(α′) Κατανομές Visitor Type Metrics Post Covid-19



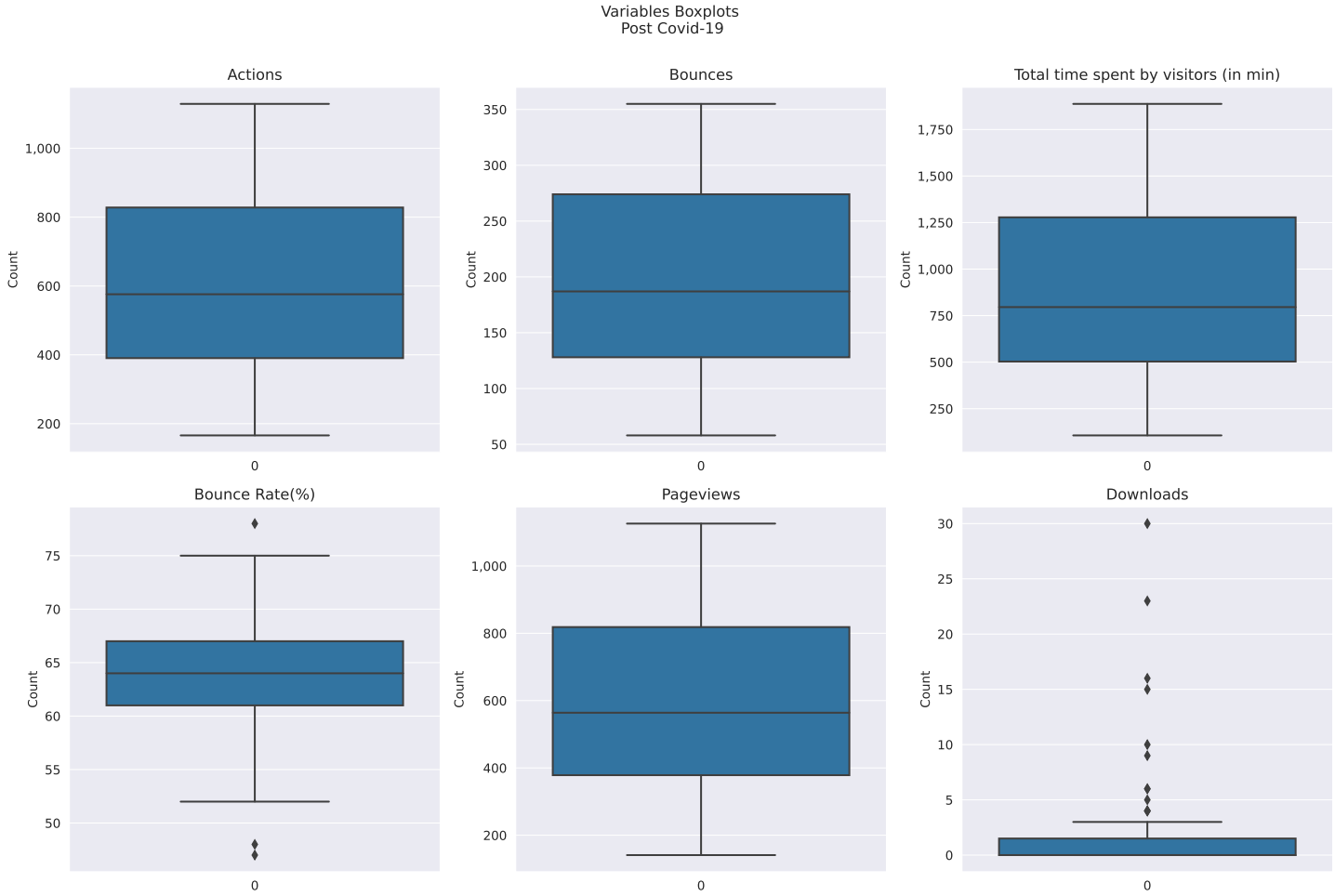
(β′) Θηκογράμματα Visitor Type Metrics Post Covid-19

	Actions	Bounces	Total time spent by visitors (in min)	Bounce Rate(%)	Pageviews	Downloads
count	67.00	67.00	67.00	67.00	67.00	67.00
mean	592.28	201.33	889.00	63.52	586.07	2.30
median	576.00	187.00	795.80	64.00	564.00	0.00
std	252.29	86.29	473.31	5.84	251.25	5.40
min	166.00	58.00	106.55	47.00	141.00	0.00
25%	390.50	128.00	503.28	61.00	378.50	0.00
75%	828.00	274.00	1,278.03	67.00	818.50	1.50
max	1,129.00	355.00	1,887.47	78.00	1,126.00	30.00

Πίνακας 5.8: Περιγραφικός πίνακας μετρικών Activity Metrics Post Covid-19



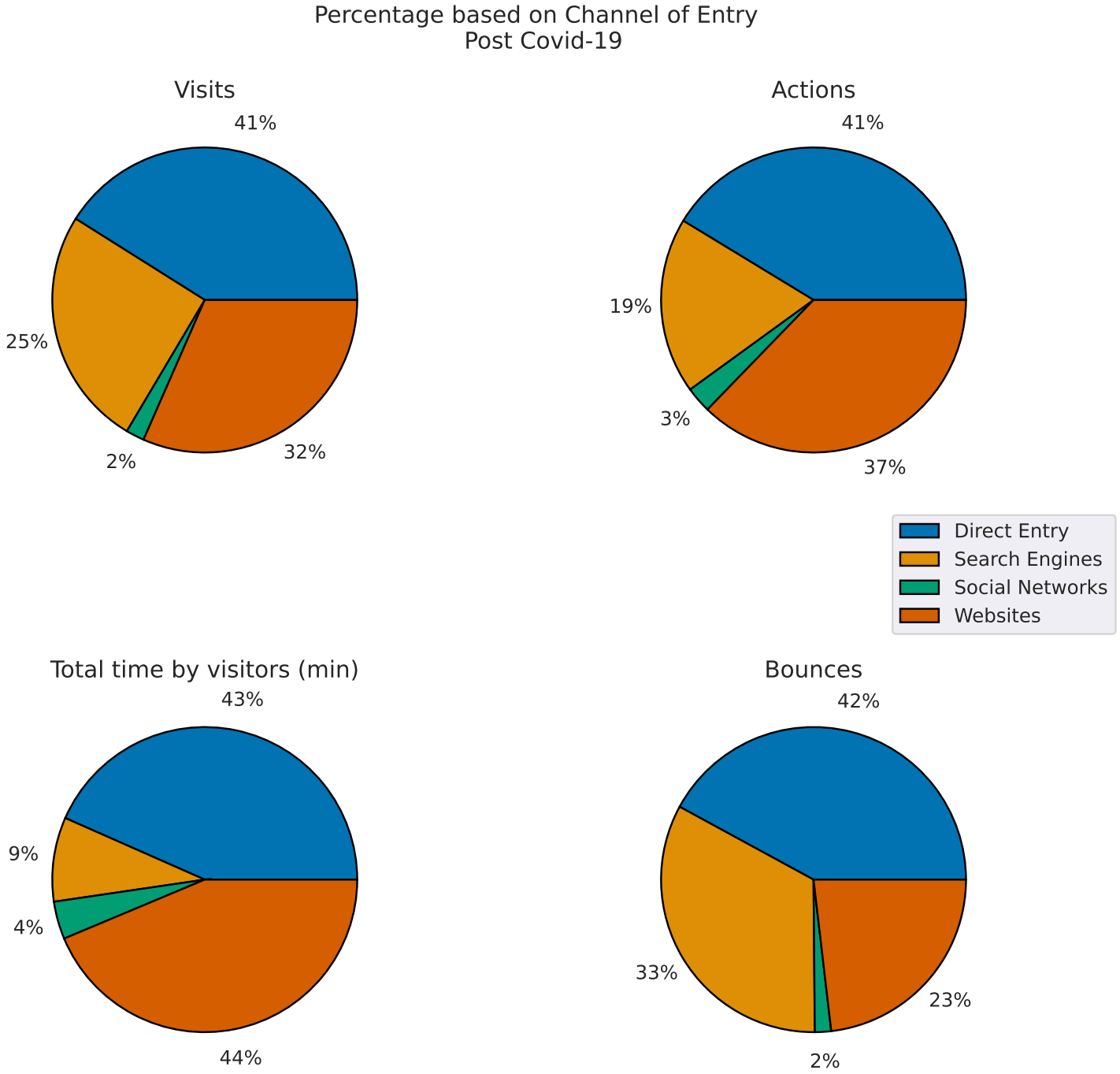
Σχήμα 5.17: Κατανομές Visitor Type Metrics Post Covid-19



Σχήμα 5.18: Θηκογράμματα Visitor Type Metrics Post Covid-19

Ομοίως για τα Activity Metrics:

- Η μεταβλητή Bounce Rate(%) τείνει να ακολουθήσει κανονική κατανομή, που όμως επηρεάζεται από τις ακραίες τιμές τις.
- Και σε αυτή την περίπτωση η μεταβλητή Downloads καταγράφει πολύ χαμηλές τιμές σε σύγκριση με τις υπόλοιπες.
- Η αναλογία Actions:Pageviews εξακολουθεί να είναι σχεδόν 1:1. Ο χρόνος παραμονής ανά επισκέπτη είναι περίπου 2.80 λεπτά.



Σχήμα 5.19: Κανάλια εισόδου Post Covid-19

Σχετικά με τα κανάλια εισόδου την περίοδο Post Covid-19, επισκέψεις μέσω Direct Entry εξακολουθούν να είναι το βασικό κανάλι, ακολουθεί η είσοδος από Websites, και καταγράφεται σημαντική αύξηση στην είσοδο από Search Engines. Η τελευταία αυτή παρατήρηση δείχνει πως υπήρξε και πάλι αλλαγή συμπεριφοράς των χρηστών, είτε λόγω εξοικείωσης τους στην αναζήτησή της πλατφόρμας μέσω μηχανών αναζήτησης, είτε λόγω κάποιας προωθητικής ενέργειας ή λόγω γενικότερης αύξησης ενδιαφέροντος προς το e-Service. Κι ενώ οι επισκέψεις ανά κανάλι εισόδου φαίνεται να παρουσιάζουν αναλογία με τα Actions, δεν

ισχύει το ίδιο για την Total Time spent by visitors (in min), δηλαδή επισκέπτες μέσω Websites παρέμειναν περισσότερη ώρα συνδεδεμένοι 44%, επισκέπτες από Direct Entry 43%, ενώ επισκέπτες από Search Engines μόλις 8% του συνολικού χρόνου από κανάλια εισόδου. Επομένως αυτοί οι επισκέπτες, αν και έκαναν κάποιες ενέργειες, ήταν πολύ συγκεκριμένες και για περιορισμένο χρόνο.

5.1.5 Συσχετίσεις & Πολυσυγγραμμικότητα στο σύνολο δεδομένων polish

Ακολουθεί έλεγχος των συσχετίσεων και ύπαρξης ή μη πολυσυγγραμμικότητας σε όλο το Data Frame polish.

Συσχετίσεις

Αναφορά στις συσχετίσεις έχει γίνει και στην Παράγραφο 4.2 καθώς ήταν απαραίτητη η κατανόησή τους για την μέθοδο επιλογής εύρεσης ελλιπών τιμών. Όπως προκύπτει από το Σχήμα 4.8 της σελίδας 44, φαίνεται να υπάρχει ισχυρή συσχέτιση μεταξύ συνόλων μεταβλητών, είτε θετική είτε αρνητική, σε όλο το εύρος του πίνακα. Παρακάτω παρατίθενται μερικές περιπτώσεις συσχέτισης του πίνακα:

Ισχυρή Θετική Συσχέτιση εμφανίζουν:

- Unique Visits $\sim$ Visits
- Actions $\sim$ Unique Pageviews
- Maximum Actions in one visit $\sim$ Pageviews

Ισχυρή - Μέτρια Αρνητική Συσχέτιση εμφανίζουν:

- Bounce Rate Returning Visits(%) $\sim$ Actions per Visit, Avg.Actions per Returning Visit

Μέτρια Συσχέτιση εμφανίζουν:

- Unique Outlinks $\sim$ Bounce Rate New Visits(%), Maxixum actions in one returning visit
- Bounce Rate(%) $\sim$ Unique Pageviews

Μηδενική ή μικρή συσχέτιση εμφανίζουν:

- Avg. Generation Time (in min) με τις περισσότερες μεταβλητές

Index	Variable	VIF
0	Visitors from Search Engines	13,805.87
1	Visitors from Social Networks	16,086.14
2	Visitors from Direct Entry	635,334.28
3	Visitors from Websites	1,913,793.03
4	Actions per Visit	76.28
5	Avg. Actions per Returning Visit	47.63
6	Avg. Actions per New Visit	15.75
7	Unique visitors	8,050.95
8	Visits	3,1242,569.58
9	Actions	8,559,443.96
10	Maximum actions in one visit	92.15
11	Bounces	531.49
12	Total time spent by visitors (in min)	334.05
13	New Visits	1,883,376.84
14	Actions by New Visits	638,364.79
15	Unique new visitors	66,046.23
16	max_actions_new	33.66
17	Pageviews	2,234,260.59
18	Unique Pageviews	952.66
19	Downloads	108.90
20	Unique Downloads	34.77
21	Outlinks	323.73
22	Unique Outlinks	68.66
23	Returning Visits	15,537,184.87
24	Actions by Returning Visits	3,077,542.36
25	Unique returning visitors	4,786.05
26	Maximum actions in one returning visit	58.32
27	Distinct websites	15.23
28	Bounce Rate(%)	119.15
29	Bounce Rate New Visits(%)	29.25
30	Bounce Rate Returning Visits(%)	46.79
31	Avg. Visit Duration (in min)	67.26
32	Avg. Duration Returning Visit (in min)	48.55
33	Avg. Duration New Visit (in min)	10.64
34	Avg. Generation Time (in min)	2.26

Πίναχας 5.9: Πίναχας VIF

Από τις παραπάνω ισχυρές συσχετίσεις μεταβλητών καθώς και από τους υπερβολικά υψηλούς δείκτες VIF συνεπάγεται πως υπάρχει σοβαρό πρόβλημα πολυσυγγραμμικότητας στα δεδομένα και εμφάνιση υψηλής γραμμικής συσχέτισης. Αυτό καθιστά ιδιαίτερα δύσκολη την ερμηνεία των συντελεστών των μεταβλητών, κυρίως σε γραμμικά μοντέλα, δηλαδή την αξιολόγηση της συνεισφοράς τους. Επιπρόσθετα, η εφαρμογή μοντέλων σε αυτό το Data Frame θα επηρεαστεί σημαντικά από την επιλογή των μεταβλητών του.

Το πρόβλημα πολυσυγγραμμικότητας μπορεί, και είναι καλό, να αντιμετωπιστεί ή τουλάχιστον να περιοριστεί, με διάφορες μεθόδους ανάλογα με τον τελικό στόχο της εκάστοτε ανάλυσης και το πεδίο εφαρμογής. Μερικοί από αυτούς είναι: αφαίρεση περιττών ή υψηλά συσχετιζομένων μεταβλητών αλλά και Recursive Feature Elimination, Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis (PCA)) για μείωση της υψηλής διάστασης του πίνακα (dimensionality reduction), μέθοδοι κανονικοποίησης όπως Lasso Regression και Ridge Regression. Επιπλέον, ορισμένα μοντέλα μηχανικής μάθησης όπως Random Forest, Gradient Boosting που βασίζονται σε δέντρα αποφάσεων ή εφαρμόζουν μεθόδους ensemble μπορούν να περιορίσουν το πρόβλημα, ωστόσο θα πρέπει να ελεγχθούν και να ερμηνευτούν με προσοχή.

Συνοψίζοντας, γραμμικά μοντέλα αντενδείκνυνται στην προκειμένη περίπτωση, και η επιλογή μεταβλητών αναμένεται να επηρεάσει την εφαρμογή στα μοντέλα που θα εφαρμοστούν.

5.1.6 Σύνοψη Αποτελεσμάτων

Από την παραπάνω διερευνητική και περιγραφική ανάλυση προκύπτει η χρήση του e-Service της Πολωνίας επηρεάστηκε σε μεγάλο βαθμό από τα γεγονότα της πανδημίας και πως κάθε περίοδος που εξετάστηκε είχε τα δικά της χαρακτηριστικά. Η περίοδος Pre Covid-19 αποτέλεσε μια πιλοτική φάση που χαρακτηρίζεται από συγκεκριμένο αριθμό επισκεπτών και γενικά πολύ χαμηλών τιμών, η Wave 1 ώθησε τη χρήση της πλατφόρμας στη μέγιστη απόδοσή της και τη μέγιστη προσέλκυση επισκεπτών, καταγράφοντας συστηματική χρήση. Το Wave 2 διατήρησε υψηλές τιμές επισκέψεων αλλά χαμηλότερα σε σχέση με το Wave 1, και τέλος για την περίοδο Post Covid-19, οι επισκέψεις μειώθηκαν σημαντικά, ωστόσο η πλατφόρμα σημείωσε συστηματικότητα στη χρήση της. Σχετικά με τις Activity Metrics, φαίνεται να υπάρχει αναλογία 1:1 μεταξύ Actions και Pageviews καθόλη τη διάρκεια της μελέτης, ενώ η μετρική Total Time spent by visitors (in min) δεν ξεπέρασε τα 7 λεπτά ανά επισκέπτη κατά μέσο όρο. Ενδιαφέρον έχει η μεταβολή των καναλιών εισόδου από Direct Entry σε Websites και η αύξηση το τελευταίο διάστημα των επισκέψεων από Search Engines.

Τέλος, σχετικά με το σύνολο δεδομένων, υπάρχουν ισχυρές συσχετίσεις (θετικές και αρνητικές) μεταξύ των μεταβλητών, και συναντάται πολύ έντονο πρόβλημα πολυσυγγραμμικότητας. Το γεγονός αυτό σε συνδυασμό με τις κατανομές των μετρικών παραπέμπει σε μη παραμετρικές μεθόδους ανάλυσης και κανονικοποιήσεις ή μοντέλα μηχανικής μάθησης που μπορούν να διαχειριστούν αυτά τα προβλήματα διάστασης.

5.2 Εφαρμογή Μοντέλων Μηχανικής Μάθησης

Σε αυτή την παράγραφο εξετάζεται ένα πρόβλημα ταξινόμησης μέσω εφαρμογής, αξιολόγησης των μοντέλων Logistic Regression, Random Forest, Gradient Boosting, XGBoost, SVM και των σημαντικότερων μεταβλητών (features). Συγκεκριμένα, χρησιμοποιείται το Data Frame polish με πλήθος παρατηρήσεων να είναι 504 κι η μεταβλητή Bounce Rate(%) ως συνάρτηση στόχου (target variable) με κλάσεις ταξινόμησης:

- Low Bounce Rate κλάση 0, για Bounce Rate(%) $< 50\%$
- High Bounce Rate κλάση 1, για Bounce Rate(%) $\geq 50\%$

Η ταξινόμηση αυτή με την επιλογής της μετρικής Bounce Rate(%), όπως και η διερεύνηση των σημαντικότερων μεταβλητών, βοηθάει στην καλύτερη κατανόηση της συμπεριφοράς των χρηστών με την αλληλεπίδραση της πλατφόρμας καθώς και τη λειτουργία αυτή καθαυτή της δεύτερης. Η επιλογή του κατωφλίου 50% έγινε βάσει συγκέντρωσης πληροφοριών σχετικών με τον μέσο όρο Bounce Rate(%) που καταγράφεται στον κλάδο της εκπαίδευσης όπως αυτή αναφέρεται στην εικόνα 2.2.

Αρχικά, τα μοντέλα που προαναφέρθηκαν εφαρμόζονται με βασικές παραμέτρους ώστε να γίνει μια πρώτη εκτίμηση της απόδοσής τους.

```
LogisticRegression(max_iter=10000, solver='saga')
RandomForestClassifier(random_state=42, bootstrap = True)
GradientBoostingClassifier(random_state=42)
XGBClassifier()
SVC(kernel='linear', probability=True)
```

Από το Data Frame polish επιλέγονται όλα τα features του, αφαιρώντας τις μεταβλητές Date, Bounce Rate New Visits(%), Bounce Rate Returning Visits(%), Bounces καθώς αυτές ενδεχομένως να επηρεάζουν υπέρμετρα την target variable και σκοπός είναι να μελετηθούν και οι υπόλοιπες μεταβλητές ως προς την σημαντικότητά τους. Ο πίνακας 5.10 της επόμενης σελίδας δείχνει την τελική επιλογή.

Index	Variable
0	Visitors from Search Engines
1	Visitors from Social Networks
2	Visitors from Direct Entry
3	Visitors from Websites
4	Actions per Visit
5	Avg. Actions per Returning Visit
6	Avg. Actions per New Visit
7	Unique visitors
8	Visits
9	Actions
10	Maximum actions in one visit
11	Total time spent by visitors (in min)
12	New Visits
13	Actions by New Visits
14	Unique new visitors
15	max_actions_new

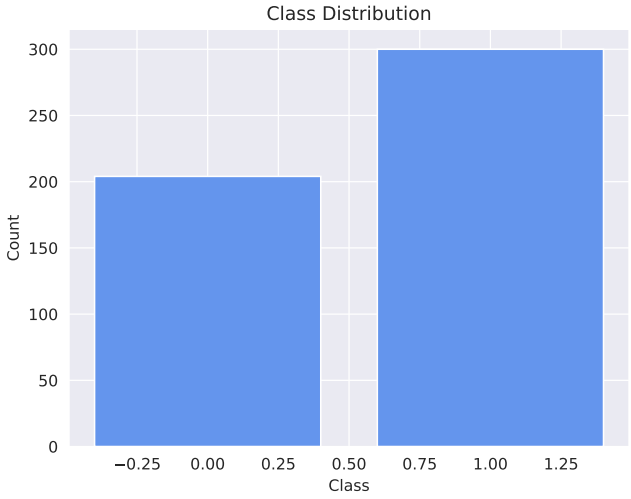
Index	Variable
16	Pageviews
17	Unique Pageviews
18	Downloads
19	Unique Downloads
20	Outlinks
21	Unique Outlinks
22	Returning Visits
23	Actions by Returning Visits
24	Unique returning visitors
25	Maximum actions in one returning visit
26	Distinct websites
27	Avg. Visit Duration (in min)
28	Avg. Duration Returning Visit (in min)
29	Avg. Duration New Visit (in min)
30	Avg. Generation Time (in min)

Πίνακας 5.10: Πίνακας features

Τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης (training set) που αποτελούν το 80% του αρχικού συνόλου και δεδομένα ελέγχου (testing set), 20% αρχικού συνόλου. Λόγω των μεγάλων διακυμάνσεων και της πολυσυγγραμμικότητας που προέκυψε από τις προηγούμενες παραγράφους, εφαρμόζεται η μέθοδος `StandardScaler` καθώς μοντέλα όπως το SVM είναι ευαίσθητα σε τέτοιους παράγοντες.

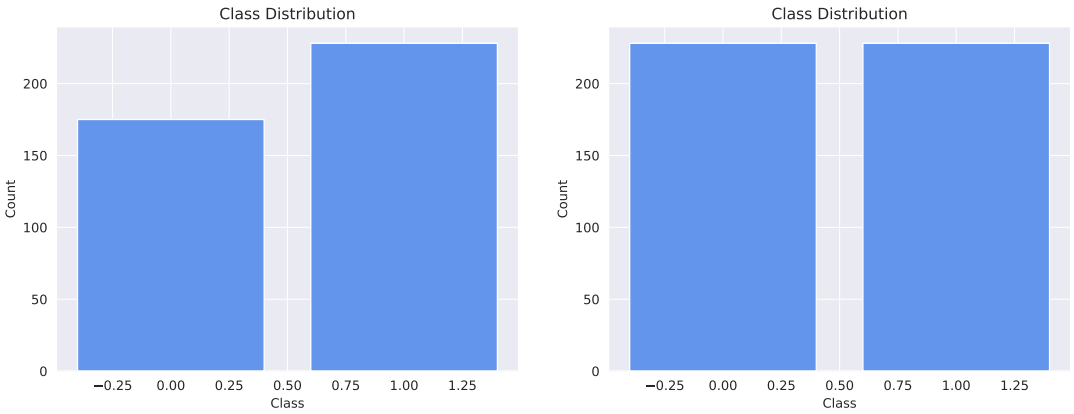
```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
↪  random_state=42)

# Scale Data
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Στη συνέχεια, από τον έλεγχο ισορροπίας των κλάσεων προκύπτει πως εμφανίζουν μικρή ασυμμετρία, όπως φαίνεται από το ακόλουθο Σχήμα 5.20.



Σχήμα 5.20: Low Bounce Rate: 204, High Bounce Rate: 300

Αν και η διαφορά δεν θεωρείται περιοριστική για την εφαρμογή των μοντέλων, εφαρμόζεται η μέθοδος Synthetic Minority Oversampling Technique (SMOTE) ώστε να εξισορροπηθούν οι κλάσεις.



(α′) Κατανομή κλάσεων χωρίς SMOTE (β′) Εφαρμογή SMOTE

Σχήμα 5.21: Κλάσεις για το y_train σύνολο

Στο πρώτο στάδιο Baseline τα μοντέλα θα εφαρμοστούν με και χωρίς τη μέθοδο SMOTE, και θα αξιολογηθεί η αναγκαιότητά της ή όχι.

5.2.1 Εφαρμογή Μοντέλων - Baseline

Στην παράγραφο αυτή παρουσιάζονται τα αποτελέσματα της εφαρμογής των μοντέλων με και χωρίς την μέθοδο SMOTE.

Classification Report μοντέλων

Logistic Regression:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.68	0.72	0.70	29
High Bounce Rate	0.89	0.86	0.87	72
Accuracy			0.82	101
Macro Avg.	0.78	0.79	0.79	101
Weighted Avg.	0.83	0.82	0.82	101

Πίνακας 5.11: Classification Report Logistic Regression

Random Forest:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.68	0.79	0.73	29
High Bounce Rate	0.91	0.85	0.88	72
Accuracy			0.83	101
Macro Avg.	0.79	0.82	0.80	101
Weighted Avg.	0.84	0.83	0.84	101

Πίνακας 5.12: Classification Report Random Forest

Gradient Boosting:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.64	0.72	0.68	29
High Bounce Rate	0.88	0.83	0.86	72
Accuracy			0.80	101
Macro Avg.	0.76	0.78	0.77	101
Weighted Avg.	0.81	0.80	0.81	101

Πίνακας 5.13: Classification Report Gradient Boosting

XGBoost:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.67	0.76	0.71	29
High Bounce Rate	0.90	0.85	0.87	72
Accuracy			0.82	101
Macro Avg.	0.78	0.80	0.79	101
Weighted Avg.	0.83	0.82	0.82	101

Πίνακας 5.14: Classification Report XGBoost

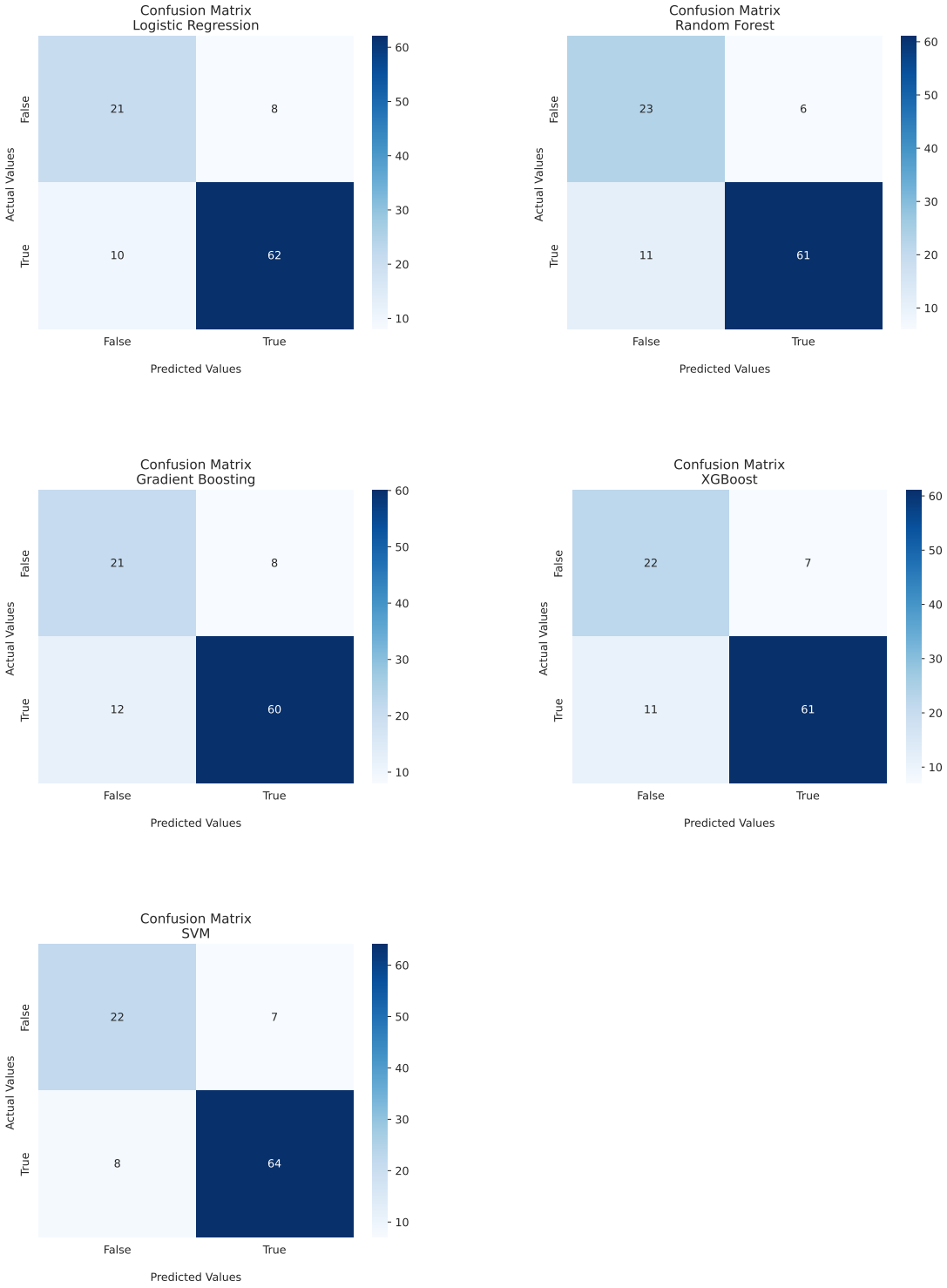
SVM:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.73	0.76	0.75	29
High Bounce Rate	0.90	0.89	0.90	72
Accuracy			0.85	101
Macro Avg.	0.82	0.82	0.82	101
Weighted Avg.	0.85	0.85	0.85	101

Πίνακας 5.15: Classification Report SVM

Όλα τα μοντέλα παρουσιάζουν αρκετά υψηλά ποσοστά ακρίβειας με το SVM να έχει το υψηλότερο ποσοστό (85%), ενώ το Gradient Boosting το χαμηλότερο (80%). Δεδομένου του μικρού συνόλου παρατηρήσεων αυτό ίσως αποτελεί ένδειξη overfitting που πρέπει να εξεταστεί. Όλα τα μοντέλα φαίνεται να προβλέπουν καλύτερα την κλάση High Bounce Rate από ό,τι την Low Bounce Rate, επηρεασμένα από την μερική ασυμμετρία των κλάσεων.

Confusion Matrices



Σχήμα 5.22: Confusion Matrix κάθε μοντέλου

Η ικανότητα των μοντέλων να προβλέπουν καλύτερα την κλάση High Bounce Rate φαίνεται και από τους πίνακες σύγχυσης (5.22) της προηγούμενης σελίδας. Το SVM μοντέλο καταφέρνει να εκτιμήσει καλύτερα τις πραγματικά θετικές τιμές TP = 64, και έχει τις λιγότερες FN = 8, σε σχέση με τα υπόλοιπα μοντέλα. Το Random Forest ήταν το πιο αποτελεσματικό στο FP = 6, ωστόσο είχε 11 λάθος TN ταξινομήσεις. Όλα τα μοντέλα έχουν περιθώρια βελτίωσης.

Performance Comparison (Training vs Testing)

	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Training Accuracy	82.13	100	99.50	100	83.62
Testing Accuracy	82.18	83.17	80.20	82.18	85.15

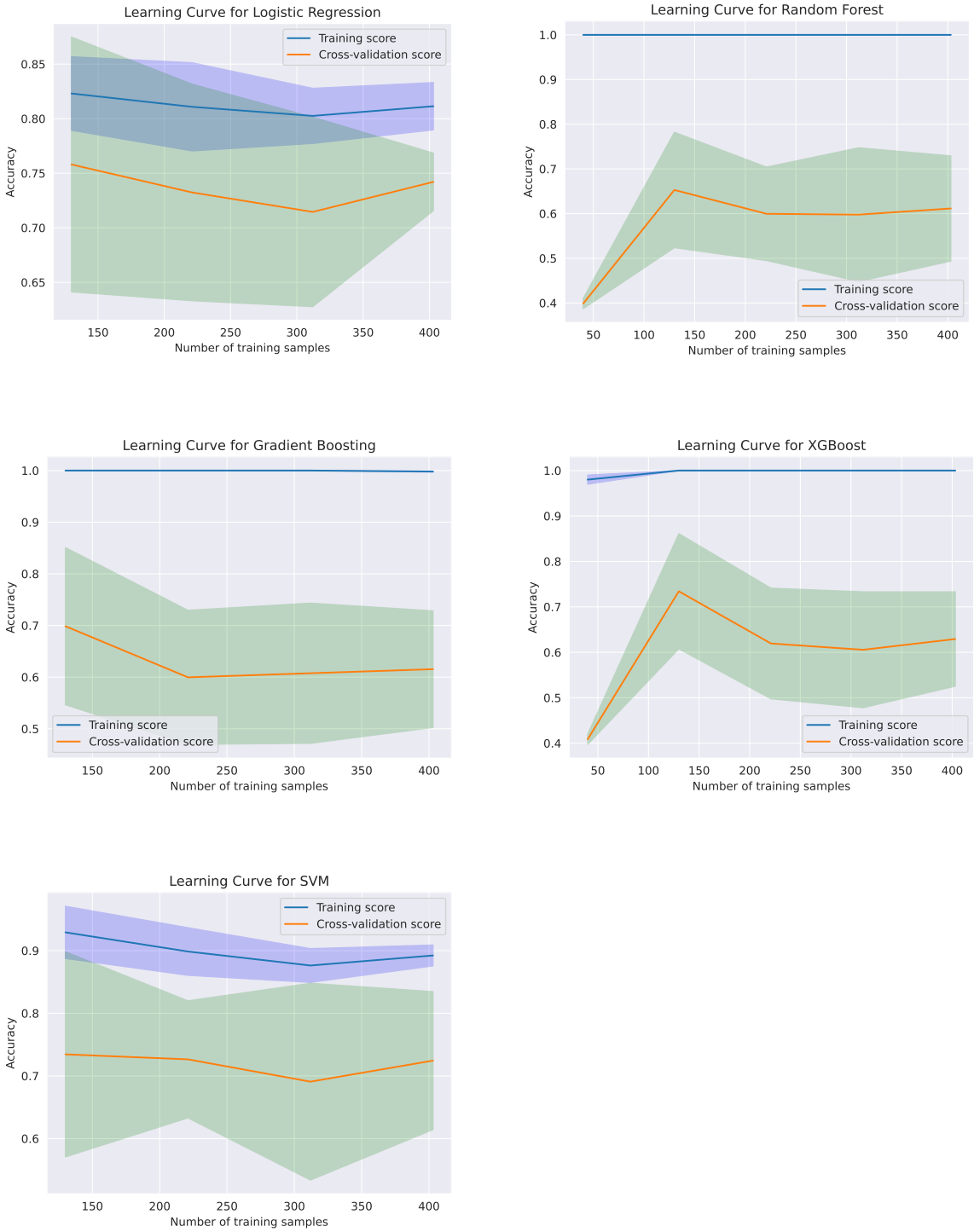
Πίνακας 5.16: Accuracy συνόλου εκπαίδευσης έναντι συνόλου ελέγχου

Η σύγκριση της ακρίβειας των training set και testing set, όπως παρουσιάζεται στον Πίνακα 5.16, επαληθεύουν πως στα μοντέλα εμφανίζεται overfitting, κυρίως για το Random Forest, κι ακολουθούν Gradient Boosting, XGBoost. Εφόσον η ακρίβεια για το training set είναι αρκετά υψηλότερη από το testing set για αυτά τα μοντέλα, σημαίνει πως ενώ στην εκπαίδευσή τους εκτιμούν σωστά τις κλάσεις, δεν μπορούν να κάνουν το ίδιο σε νέα δεδομένα. Αξιοσημείωτο είναι πως για τα μοντέλα Logistic Regression και SVM, παρατηρείται πως το training set και το testing set βρίσκονται αρκετά κοντά, για την ακρίβεια το testing set είναι υψηλότερο.

Παρατηρώντας τις καμπύλες εκπαίδευσης της επόμενης σελίδας (σελ. 81), το παραπάνω είναι προφανές, δηλαδή, πως όλα τα μοντέλα εμφανίζουν overfitting, με Logistic Regression και SVM να προσαρμόζονται καλύτερα. Για τη συμπεριφορά αυτή των δυο τελευταίων μοντέλων θα πρέπει να ληφθεί υπόψη πως και τα δύο αποτελούν γραμμικά μοντέλα, και πως ο τρόπος λειτουργίας τους επηρεάζεται από τους συντελεστές των μεταβλητών και από την πολυσυγγραμμικότητα των δεδομένων. Έτσι, η πολυπλοκότητα και το μικρό σύνολο παρατηρήσεων μπορεί να ενισχύει το παραπάνω αποτέλεσμα.

Για την αντιμετώπιση του overfitting μπορούν να πραγματοποιηθούν διάφορες ενέργειες, όπως εφαρμογή σε περισσότερα δεδομένα, διαφορετικές μέθοδοι κανονικοποίησης και εξασφάλιση συμμετρίας των κλάσεων.

Learning Curves



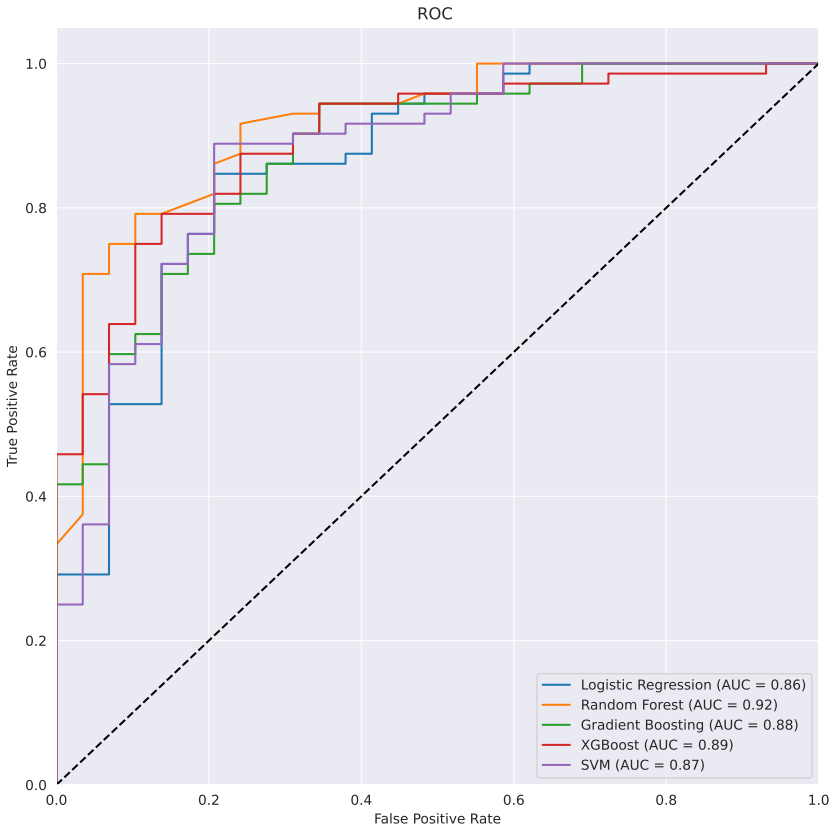
Σχήμα 5.23: Οι καμπύλες εκμάθησης δείχνουν την υπερεκπαίδευση των μοντέλων

	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Accuracy	82.10%	83.17%	80.20%	82.18%	85.15%
Precision	0.8857	0.9104	0.8824	0.8971	0.9014
Recall	0.8611	0.8472	0.8333	0.8472	0.8889
F1-score	0.8732	0.8777	0.8571	0.8714	0.8951
ROC	0.8625	0.9159	0.8774	0.8937	0.8745

Πίνακας 5.17: Μετρικές αξιολόγησης μοντέλων

Τα αποτελέσματα και οι μετρικές των μοντέλων συνοψίζονται στον Πίνακα 5.17.

- SVM - επιτυγχάνει τη μεγαλύτερη ακρίβεια (85%)
- Random Forest - υψηλότερο Precision και ROC, δηλαδή καλύτερος διαχωρισμός μεταξύ κλάσεων
- Gradient Boosting - χαμηλότερες αξιολογήσεις σε όλες τις μετρικές μεταξύ των μοντέλων
- σε όλα τα μοντέλα παρατηρείται overfitting, κυρίως στο Random Forest



Σχήμα 5.24: Καμπύλη ROC

K-Fold Cross Validation

	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Fold 1	77.78	80.25	81.48	85.19	72.84
Fold 2	80.25	79.01	75.31	82.72	79.01
Fold 3	85.19	86.42	82.71	82.72	82.71
Fold 4	81.25	80.00	77.50	78.75	85.00
Fold 5	75.00	86.25	85.00	85.00	76.25
Mean Score Accuracy (%)	**79.89**	**82.39**	**80.40**	**82.87**	**79.16**

Πίνακας 5.18: K-Fold Cross Validation με k= 5

Το XGBoost αποδίδει την υψηλότερη τιμή 82.87% και ακολουθεί το Random Forest με 82.39%. Τα δύο μοντέλα παρουσιάζουν μικρή διακύμανση και σταθερότητα μεταξύ των Fold. Ακολουθεί το Gradient Boosting με 80.40% και τέλος Logistic Regression με 79.89% και SVM με 79.16%.

Συνοπτικά, τα μοντέλα που βασίζονται σε μεθόδους ensemble αποδίδουν καλύτερα και παρουσιάζεται ομοιογένεια στα αποτελέσματα των διαχωρισμών. Αντίθετα, τα μοντέλα που βασίζονται σε γραμμικές σχέσεις μεταξύ μεταβλητών, φαίνεται να επηρεάζονται από τα δεδομένα και έχουν μικρότερο μέσο όρο ακρίβειας και διακύμανση ανά K-Fold Cross Validation.

Τα αποτελέσματα αυτά δείχνουν πως υπάρχουν περιθώρια βελτίωσης μέσω κατάλληλων επεμβάσεων, όπως παραμετροποιήσεις των μοντέλων ή περαιτέρω κανονικοποίηση των δεδομένων και διαχείριση ασυμμετριών.

Classification Report με τη μέθοδο SMOTE

Logistic Regression:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.67	0.76	0.71	29
High Bounce Rate	0.90	0.85	0.87	72
Accuracy			0.82	101
Macro Avg.	0.78	0.80	0.79	101
Weighted Avg.	0.83	0.82	0.82	101

Πίνακας 5.19: Classification Report Logistic Regression με SMOTE

Random Forest:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.69	0.76	0.72	29
High Bounce Rate	0.90	0.86	0.88	72
Accuracy			0.83	101
Macro Avg.	0.79	0.81	0.80	101
Weighted Avg.	0.84	0.83	0.83	101

Πίνακας 5.20: Classification Report Random Forest με SMOTE

Gradient Boosting:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.66	0.72	0.69	29
High Bounce Rate	0.88	0.85	0.87	72
Accuracy			0.81	101
Macro Avg.	0.77	0.79	0.78	101
Weighted Avg.	0.82	0.81	0.81	101

Πίνακας 5.21: Classification Report Gradient Boosting με SMOTE

XGBoost:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.63	0.76	0.69	29
High Bounce Rate	0.89	0.82	0.86	72
Accuracy			0.80	101
Macro Avg.	0.76	0.79	0.77	101
Weighted Avg.	0.82	0.80	0.81	101

Πίνακας 5.22: Classification Report XGBoost με SMOTE

SVM:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.68	0.79	0.73	29
High Bounce Rate	0.91	0.85	0.88	72
Accuracy			0.83	101
Macro Avg.	0.79	0.82	0.80	101
Weighted Avg.	0.84	0.83	0.84	101

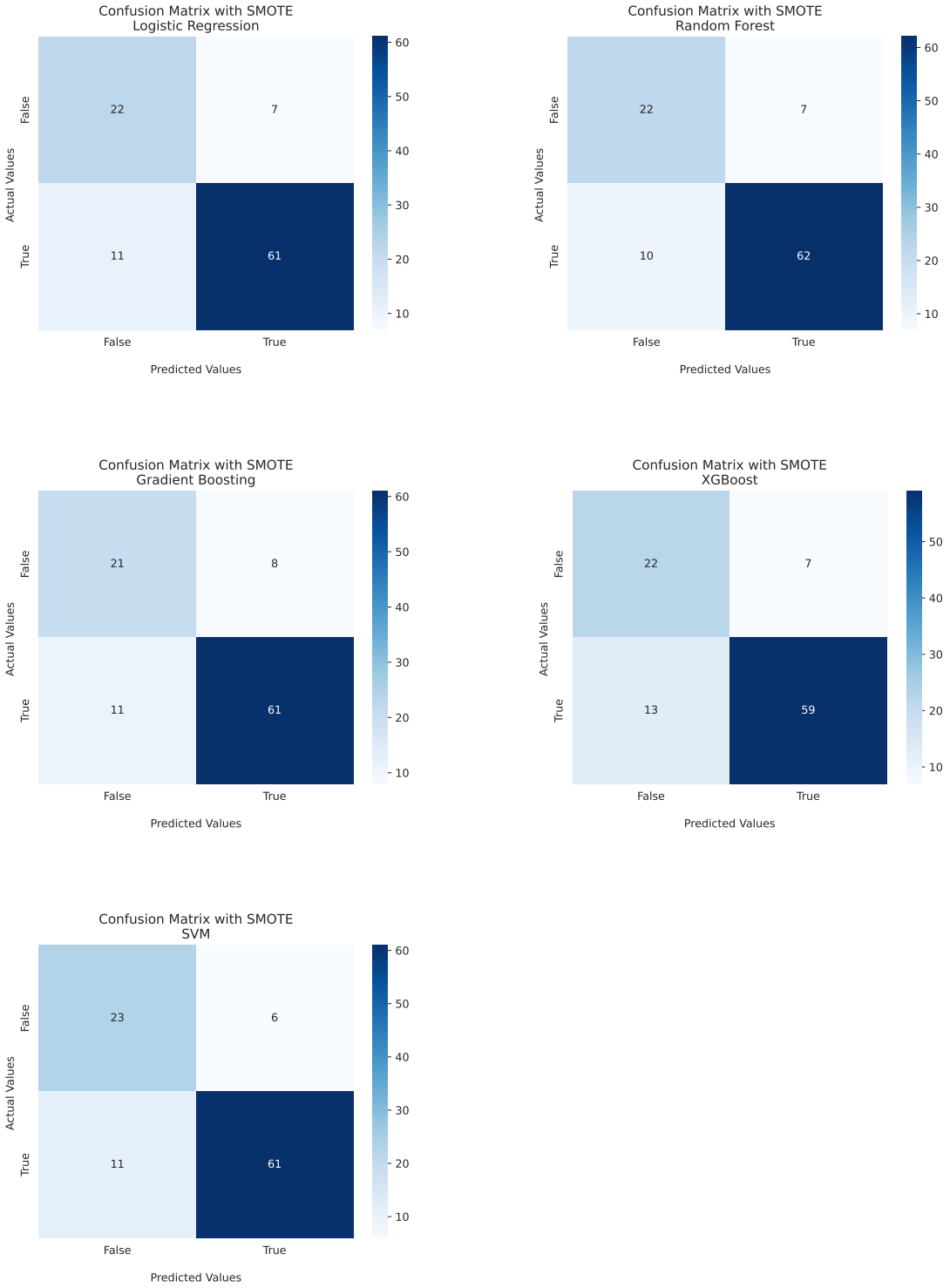
Πίνακας 5.23: Classification Report SVM με SMOTE

Η εφαρμογή της μεθόδου SMOTE δεν επηρέασε τα μοντέλα σε σημαντικό βαθμό ως προς το Accuracy.

Οι αλλαγές στο Accuracy σημειώνονται ως εξής:

- SVM: 85% → 83%
- XGBoost: 82% → 80%
- Gradient Boosting: 80% → 81%
- Random Forest και Logistic Regression διατήρησαν τις τιμές τους

Confusion Matrices με μέθοδο SMOTE



Σχήμα 5.25: Confusion Matrix κάθε μοντέλου μετά την εφαρμογή SMOTE

Από τους πίνακες σύγχυσης 5.25 φαίνονται πως τα μοντέλα δεν είχαν ουσιαστικές αλλαγές στην απόδοση ταξινόμησης. Το μοντέλο XGBoost και SVM επηρεάστηκαν περισσότερο εμφανίζοντας μείωση στην εκτίμηση των πραγματικά θετικών TP παρατηρήσεων και αύξηση στον αριθμό των FN. Δηλαδή, η μέθοδος SMOTE, στην προσπάθεια εξισορρόπησης των κλάσεων, ίσως εισήγαγε κάποιο θόρυβο που επηρέασε τα δύο μοντέλα.

Performance Comparison (Training vs Testing) - SMOTE					
	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Training Accuracy	82.24	100.0	99.56	100.0	84.65
Testing Accuracy	82.18	83.17	81.19	80.20	83.17

Πίνακας 5.24: Accuracy συνόλου εκπαίδευσης έναντι συνόλου ελέγχου - με SMOTE

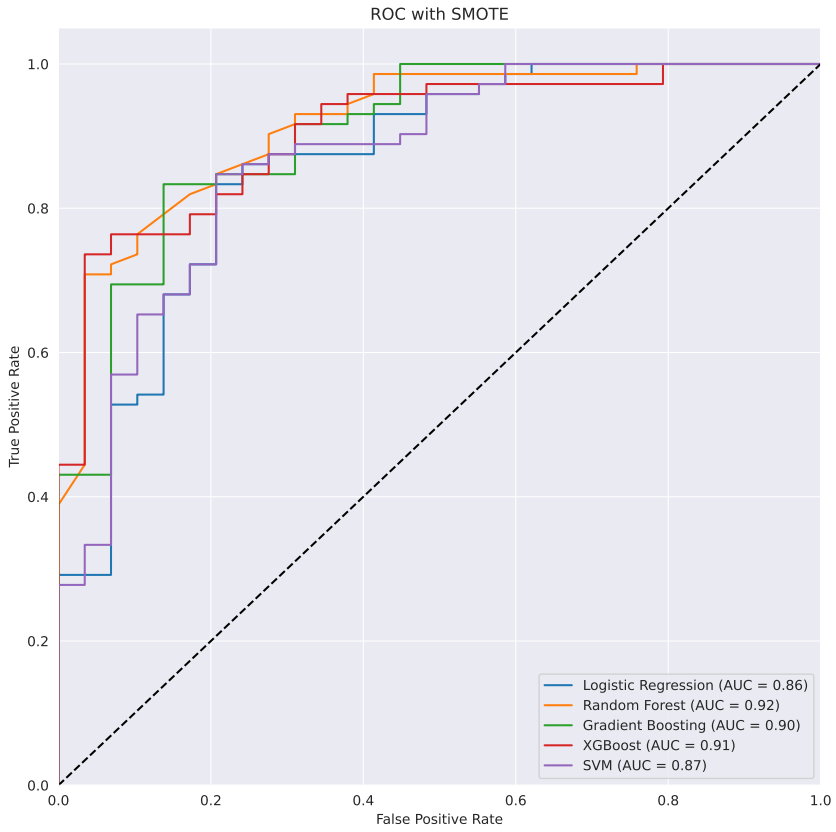
Συγκρίνοντας το training set και το testing set, τα μοντέλα εμφανίζουν overfitting και σε αυτή την περίπτωση με μικρές αλλαγές, κυρίως για το XGBoost και το SVM.

	Μετρικές αξιολόγησης μετά την εφαρμογή SMOTE				
	Logistic Regression SMOTE	Random Forest SMOTE	Gradient Boosting SMOTE	XGBoost SMOTE	SVM SMOTE
Accuracy	82.18%	83.17%	81.19%	80.20%	83.17%
Precision	0.8971	0.8986	0.8841	0.8939	0.9104
Recall	0.8472	0.8611	0.8472	0.8194	0.8472
F1-score	0.8714	0.8794	0.8652	0.8551	0.8777
ROC	0.8616	0.9162	0.9023	0.9066	0.8678

Πίνακας 5.25: Μετρικές αξιολόγησης και σύγκρισης μοντέλων με τη μέθοδο SMOTE

Από τον συγκεντρωτικό πίνακα των μετρικών αξιολόγησης προκύπτουν τα εξής:

- Random Forest και SVM αποδίδουν καλύτερα με ακρίβεια 83.17% σε σχέση με τα υπόλοιπα μοντέλα
- Gradient Boosting και XGBoost είχαν καλύτερα αποτελέσματα για την καμπύλη ROC κατά 2%. Αυτό σημαίνει πως βελτιώθηκαν στον διαχωρισμό τον κλάσεων
- οι μετρικές Precision, Recall και F1-score είτε παρέμειναν σταθερές είτε δεν μεταβλήθηκαν σε σημαντικό βαθμό
- η Recall επηρεάστηκε περισσότερο από την εφαρμογή της μεθόδου SMOTE σε σχέση με τις υπόλοιπες μετρικές



Σχήμα 5.26: ROC μετά την εφαρμογή της μεθόδου SMOTE

K-Fold Cross Validation με SMOTE

	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Fold 1	79.35	81.52	81.52	82.61	82.61
Fold 2	79.12	78.02	72.53	78.02	81.31
Fold 3	85.71	89.01	84.62	85.71	86.81
Fold 4	73.62	84.62	84.62	86.81	73.62
Fold 5	84.62	90.11	90.11	89.01	86.81
Mean Score Accuracy (%)	**80.48**	**84.66**	**82.68**	**84.43**	**82.24**

Πίνακας 5.26: K-Fold Cross Validation με k= 5 με SMOTE

Παρατηρείται μικρή αύξηση του μέσου όρου ακρίβειας σε όλα τα μοντέλα. Και αυτή τη φορά Random Forest και XGBoost επιτυγχάνουν τους υψηλότερους μέσους όρους ακρίβειας με 84.66% και 84.48% αντίστοιχα, με μικρή διακύμανση μεταξύ των τιμών των Fold. Ακολουθούν τα υπόλοιπα μοντέλα που ωστόσο εμφανίζουν ελαφρώς μεγαλύτερη διακύμανση.

5.2.1.1 Σύνοψη αποτελεσμάτων και σύγκριση μεθόδων

Συγκρίνοντας τις εφαρμογές των μοντέλων χωρίς SMOTE και με χρήση της μεθόδου, προκύπτει πως αυτή η μέθοδος oversampling δεν επηρέασε σημαντικά την απόδοση των μοντέλων. Και στις δύο περιπτώσεις στα μοντέλα υπήρξε overfitting που ενδεχομένως να οφείλεται στο μικρό πλήθος παρατηρήσεων, την κατανομή αυτών, την ύπαρξη ακραίων, αλλά δικαιολογημένων τιμών, καθώς και την πολυσυγγραμμικότητα, που άλλα μοντέλα διαχειρίζονται επαρκώς άλλα όχι.

Υψηλότερη ακρίβεια εμφάνισαν τα μοντέλα XGBoost (82.87%) και Random Forest (82.39%) μετά τον έλεγχο K-Fold Cross Validation, ενώ το μοντέλο που πέτυχε ορθότερες ταξινομήσεις είναι το Random Forest και στις δύο περιπτώσεις.

Αξίζει, ακόμη, να σημειωθεί η διαφορά απόδοσης μεταξύ των γραμμικών μοντέλων Logistic Regression, SVM και των ensemble μοντέλων (Random Forest, Gradient Boosting, XGBoost). Τα πρώτα ανταποκρίθηκαν ικανοποιητικά στο μικρό πλήθος δεδομένων, ωστόσο τα μοντέλα ensemble είχαν, σχετικά, καλύτερες επιδόσεις και στις δύο περιπτώσεις εφαρμογής. Το Random Forest και το XGBoost ήταν μοντέλα με την συνολικά καλύτερη απόδοση.

Είναι προφανές πως η πολυπλοκότητα των μοντέλων και ο διαφορετικός τρόπος λειτουργίας τους επηρεάζει το ποσοστό ακρίβειας και σε συνδυασμό με το μικρό πλήθος παρατηρήσεων προκύπτει overfitting και υψηλές φαινομενικά αποδόσεις. Εφόσον όμως σκοπός της εφαρμογής είναι η κατανόηση των μοντέλων και της συνεισφοράς των επεξηγηματικών μεταβλητών ως προς την κατάταξη της μετρικής Bounce Rate(%), έχει νόημα η εφαρμογή λιγότερο πολύπλοκων μοντέλων που, ενδεχομένως, να οδηγήσουν σε χαμηλότερο Accuracy. Υπό την έννοια της επεξήγησης explainability, μία τέτοια προσέγγιση θεωρείται αποδεκτή.

5.2.2 Εφαρμογή με Recursive Feature Elimination

Για την αντιμετώπιση της πολυπλοκότητας του Data Frame polish και της αναμενόμενης επιρροής στα μοντέλα ταξινόμησης, εφαρμόζεται η μέθοδος Recursive Feature Elimination (RFE). Με την απλοποίηση που πραγματοποιείται μέσω της διαγραφής των λιγότερο σημαντικών μεταβλητών για κάθε μοντέλο, επιχειρείται τόσο η βελτίωση της απόδοσής τους όσο και η μια πιο ακριβής και κατανοητή ερμηνεία των αποτελεσμάτων τους.

Στην συγκεκριμένη περίπτωση επιλέγεται ο αριθμός των τελικών features να είναι 10. Για κάθε μοντέλο που εφαρμόζεται, απαλείφονται διαδοχικά μεταβλητές έως ότου φτάσουν τον αριθμό αυτό.

```
selected_features = []

for name, model in models:
  # Initialise RFE for each model, select 10 of features
  rfe = RFE(estimator=model, n_features_to_select=10)

  rfe.fit(X_train, y_train)

  # Get selected features
  selected_features = rfe.support_ # returns TRUE = feature selected, FALSE =
  ↪  feature not selected

  X_train_rfe = X_train[:, selected_features]
  X_test_rfe = X_test[:, selected_features]

  model.fit(X_train_rfe, y_train)
```

Όπως προέκυψε από την εφαρμογή των μοντέλων στην Baseline περίπτωση, με ή χωρίς SMOTE, η μέθοδος SMOTE δεν συνεισέφερε σημαντικά στην βελτίωση των μοντέλων. Επομένως, για ελαχιστοποίηση τυχόν θορύβου που μπορεί να προκαλέσει το oversampling στα επόμενα βήματα εφαρμογής RFE και επανεφαρμογής των μοντέλων, χρησιμοποιούνται τα μοντέλα με τις αρχικές παραμετρικές χωρίς SMOTE.

Classification Report μοντέλων μετά την εφαρμογή RFE

Logistic Regression:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.65	0.52	0.58	29
High Bounce Rate	0.82	0.89	0.85	72
Accuracy			0.78	101
Macro Avg.	0.74	0.70	0.72	101
Weighted Avg.	0.77	0.78	0.77	101

Πίνακας 5.27: Logistic Regression με RFE

Random Forest:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.65	0.83	0.73	29
High Bounce Rate	0.92	0.82	0.87	72
Accuracy			0.82	101
Macro Avg.	0.79	0.82	0.80	101
Weighted Avg.	0.84	0.82	0.83	101

Πίνακας 5.28: Random Forest με RFE

Gradient Boosting:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.71	0.76	0.73	29
High Bounce Rate	0.90	0.88	0.89	72
Accuracy			0.84	101
Macro Avg.	0.80	0.82	0.81	101
Weighted Avg.	0.85	0.84	0.84	101

Πίνακας 5.29: Gradient Boosting με RFE

XGBoost:

	Precision	Recall	F1-score	Support
Low Bounce Rate	0.64	0.79	0.71	29
High Bounce Rate	0.91	0.82	0.86	72
Accuracy			0.81	101
Macro Avg.	0.77	0.81	0.78	101
Weighted Avg.	0.83	0.81	0.82	101

Πίνακας 5.30: XGBoost με RFE

SVM:

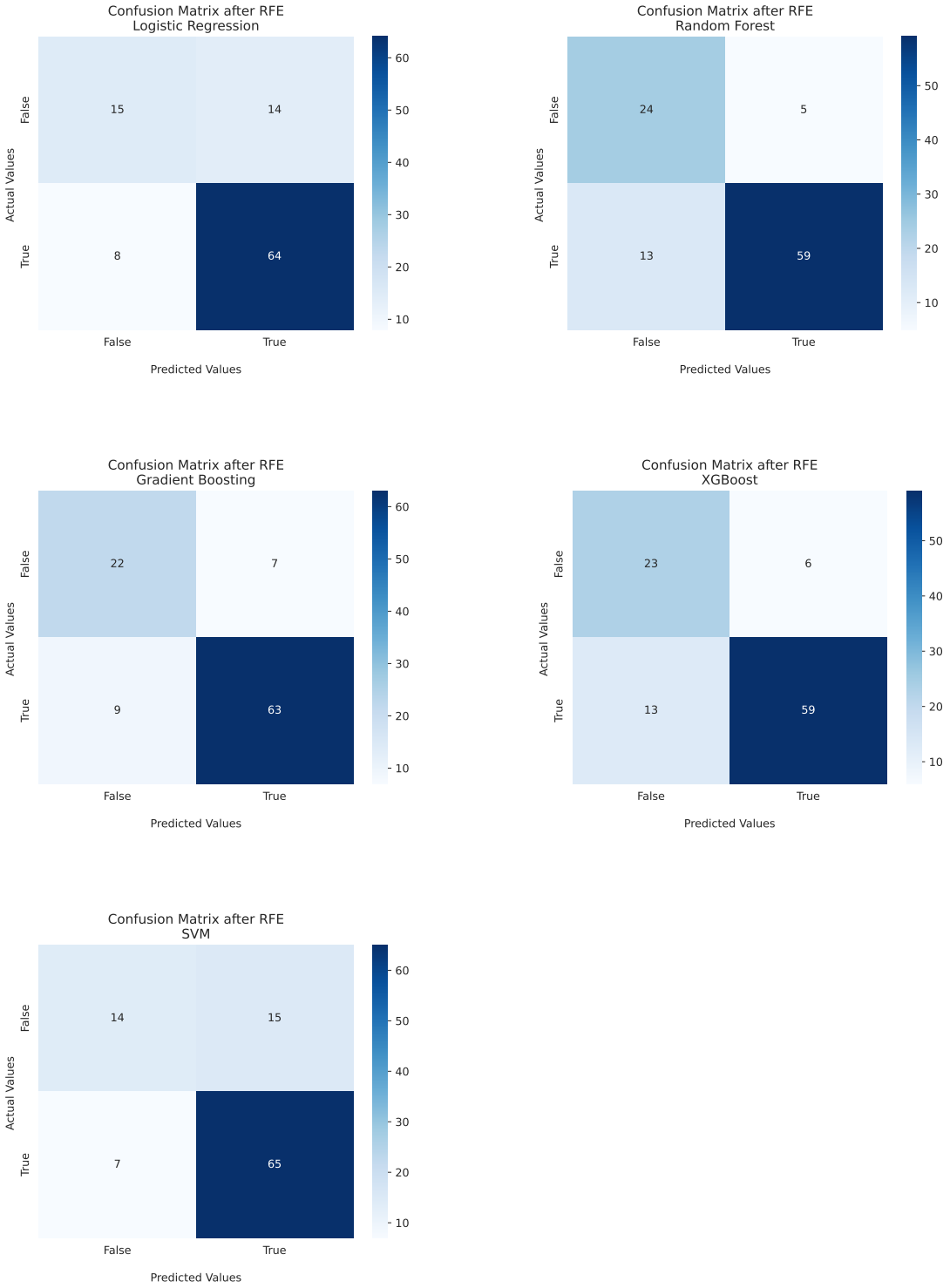
	Precision	Recall	F1-score	Support
Low Bounce Rate	0.67	0.48	0.56	29
High Bounce Rate	0.81	0.90	0.86	72
Accuracy			0.78	101
Macro Avg.	0.74	0.69	0.71	101
Weighted Avg.	0.77	0.78	0.77	101

Πίνακας 5.31: SVM με RFE

Παρατηρούνται οι εξής αλλαγές ως προς το Accuracy σε σύγκριση με την Baseline εφαρμογή

- Logistic Regression 82% → 78%
- Random Forest 83% → 82%
- Gradient Boosting 80% → 84%
- XGBoost 82% → 81%
- SVM 85% → 78%

Confusion Matrices μετά την εφαρμογή RFE



Σχήμα 5.27: Confusion Matrix κάθε μοντέλου μετά την εφαρμογή RFE

Η εφαρμογή της RFE φαίνεται να επηρέασε τον τρόπο με τον οποίο προβλέπουν τις ταξινομήσεις τα μοντέλα. Η επιρροή είναι πιο έντονη στα μοντέλα Logistic Regression, SVM και Gradient Boosting που σημείωσαν αύξηση των TP. Για τα Logistic Regression και SVM ωστόσο αυξήθηκαν σημαντικά και τα FP που σημαίνει πως τα μοντέλα δυσκολεύτηκαν να ξεχωρίσουν προβλέψεις της κλάσης Low Bounce Rate και επηρεάστηκαν από τον θόρυβο στα δεδομένα. Ο αντίστοιχος αριθμός των FP για Random Forest και XGBoost μειώθηκε κατά μία μονάδα αλλά επίσης μειώθηκε και ο αριθμός των TP.

Performance Comparison with RFE

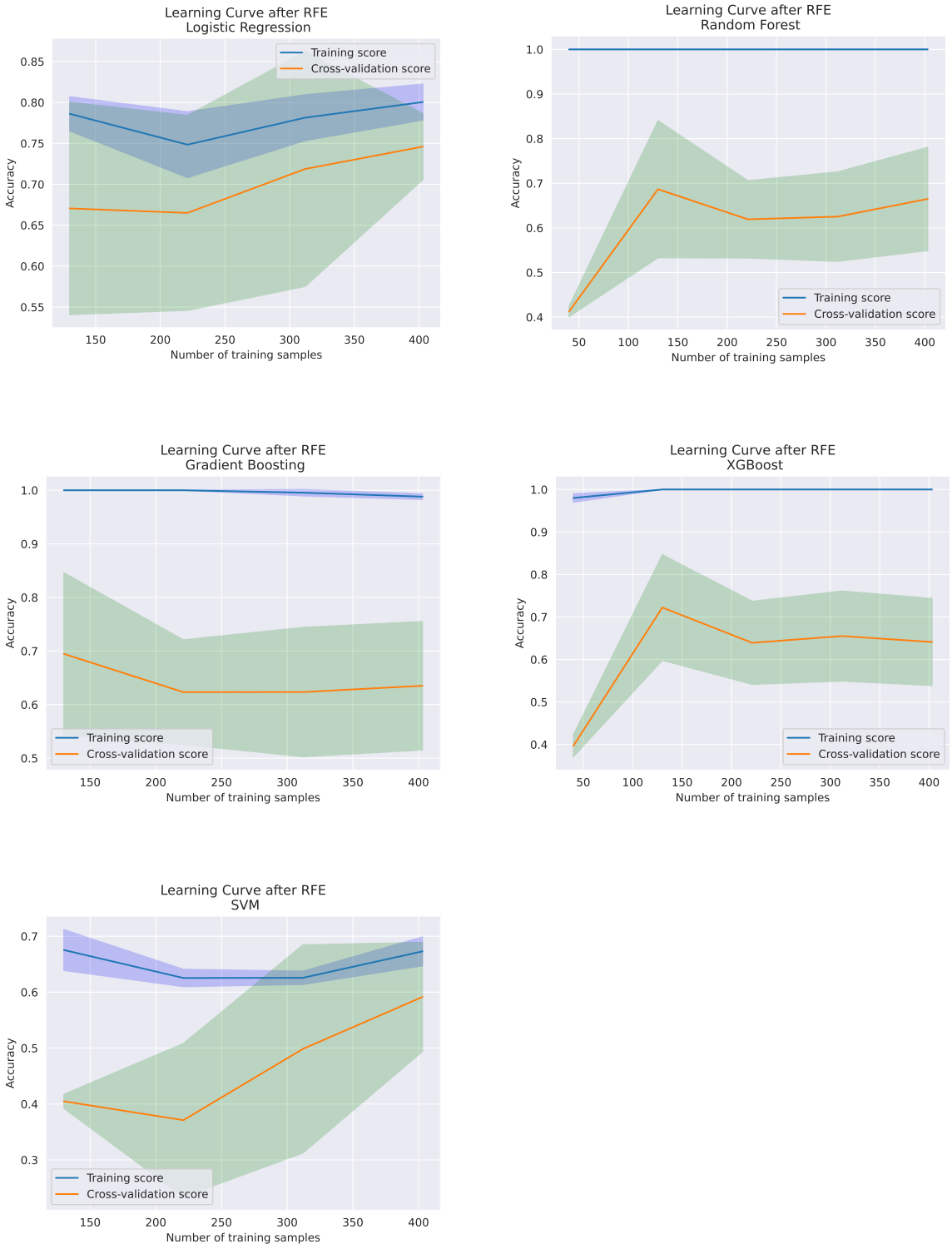
	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Training Accuracy	79.65	100	100	100	79.00
Testing Accuracy	78.21	82.18	84.16	81.19	78.22

Πίνακας 5.32: Accuracy συνόλου εκπαίδευσης έναντι συνόλου ελέγχου με RFE

Συγκρίνοντας το training set και το testing set κάθε μοντέλου φαίνεται πως τα μοντέλα παρουσιάζουν και πάλι overfitting. Random Forest, Gradient Boosting, XGBoost συγκεκριμένα αποδίδουν 100% στο training set ενώ έχουν μεγάλη διαφορά με την ακρίβεια στο testing set. Τα Logistic Regression και SVM δίνουν μικρότερα ποσοστά ακρίβειας σε σχέση με τη Baseline εφαρμογή τόσο για το training set όσο και για το testing set. Αν και η διαφορά μεταξύ training set και testing set είναι μικρότερη για αυτά τα δύο μοντέλα, και αυτά φαίνεται να εμφανίζουν overfitting. Από τα παραπάνω, η εφαρμογή της μεθόδου RFE φαίνεται να μείωσε την ακρίβεια των μοντέλων, με εξαίρεση το Gradient Boosting που σημείωσε αύξηση, όμως δεν άλλαξε το overfitting.

Ενδιαφέρον έχουν οι καμπύλες μάθησης και η σύγκρισή τους με της Baseline κατάστασης (σελ. 81). Τα γραφήματα δείχνουν πως σε όλα τα μοντέλα υπάρχει overfitting. Το Random Forest και το XGBoost δεν παρουσιάζουν σημαντικές αλλαγές με τη Baseline εφαρμογή, ενώ τα γραφήματα των δύο γραμμικών μοντέλων, Logistic Regression και SVM, δείχνουν να έχουν μεγάλη διακύμανση στην εφαρμογή τους και οι καμπύλες έχουν τάση για σύγκλιση σε τιμές γύρω στο 80%. Όπως και πριν τα δύο αυτά μοντέλα φαίνεται να επηρεάζονται άμεσα από την αφαίρεση μεταβλητών και να είναι πιο ευαίσθητα στις αλλαγές που επιδέχονται οι συντελεστές τους.

Learning Curves μετά την εφαρμογή RFE

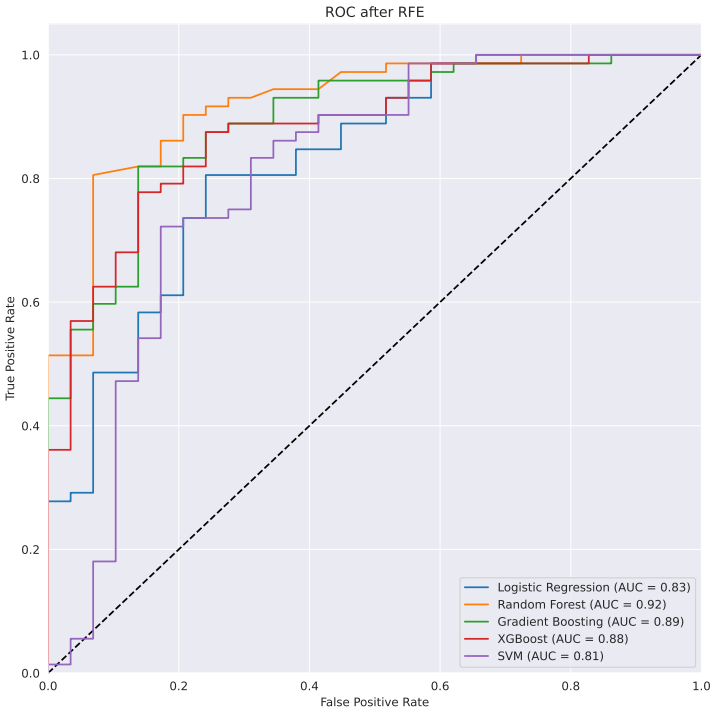


Σχήμα 5.28: Οι καμπύλες εκμάθησης μετά την εφαρμογή του RFE

	Μετρικές αξιολόγησης μετά την εφαρμογή RFE				
	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Accuracy	78.22%	82.18%	84.16%	81.19%	78.22%
Precision	0.8205	0.9219	0.9000	0.9077	0.8125
Recall	0.8889	0.8194	0.8750	0.8194	0.9028
F1-score	0.8533	0.8676	0.8873	0.8613	0.8553
ROC	0.8276	0.9217	0.8922	0.8812	0.8060

Πίνακας 5.33: Μετρικές αξιολόγησης και σύγκρισης μοντέλων μετά την αφαίρεση μεταβλητών με RFE

- Logistic Regression και SVM: επηρεάστηκαν σημαντικά περισσότερο από τα υπόλοιπα μοντέλα σε όλες τις μετρικές.
- Gradient Boosting: ανταποκρίθηκε καλύτερα στην αφαίρεση μεταβλητών επιτυγχάνοντας το υψηλότερο Accuracy.
- Random Forest και XGBoost: είχαν μικρή μείωση στο Recall



Σχήμα 5.29: ROC μετά την εφαρμογή RFE

K-Fold Cross Validation μετά την εφαρμογή RFE

	Logistic Regression	Random Forest	Gradient Boosting	XGBoost	SVM
Fold 1	76.54	76.54	79.01	81.48	76.54
Fold 2	85.19	77.78	76.54	77.78	83.95
Fold 3	80.25	81.48	85.19	83.95	79.01
Fold 4	78.75	81.25	80.00	81.25	78.75
Fold 5	67.50	83.75	80.00	87.50	73.75
Mean Score Accuracy (%)	**77.65**	**80.16**	**80.15**	**82.39**	**78.49**

Πίνακας 5.34: K-Fold Cross Validation με k= 5 μετά την εφαρμογή RFE

Εφαρμόζοντας K-Fold Cross Validation παρατηρείται πως ο μέσος όρος ακρίβειας όλων των μοντέλων σημείωσε μικρή πτώση. Επομένως η αφαίρεση μεταβλητών οδήγησε σε μερική απώλεια πληροφορίας. Από αυτή την απώλεια το Gradient Boosting και το XGBoost επηρεάστηκαν το λιγότερο αφού η ακρίβεια μειώθηκε κατά 0.025% για το πρώτο και 0.048% για το δεύτερο. Όλα τα ensemble μοντέλα έχουν σχετικά μικρές διακυμάνσεις μεταξύ των Fold ενώ το Logistic Regression, SVM εξακολουθούν να έχουν συγκριτικά μεγαλύτερες διακυμάνσεις.

5.2.2.1 Συμπεράσματα

Αν και στόχος της μεθόδου RFE είναι η βελτιστοποίηση των μοντέλων με την αφαίρεση λιγότερο σημαντικών μεταβλητών, φαίνεται πως σε αυτή την περίπτωση δεν ωφέλησε στην ακρίβεια των μοντέλων. Η απαλοιφή μεταβλητών οδήγησε σε μοντέλα λιγότερο πολύπλοκα που με μικρότερη απόδοση από πριν. Υψηλότερη ακρίβεια πέτυχε το Gradient Boosting ενώ καλύτερη ικανότητα ταξινόμησης διατήρησε το Random Forest. Τα γραμμικά μοντέλα Logistic Regression και SVM επηρεάστηκαν περισσότερο από την εφαρμογή της απαλοιφής. Τέλος, το XGBoost είχε καλύτερα αποτελέσματα στον K-Fold Cross Validation έλεγχο.

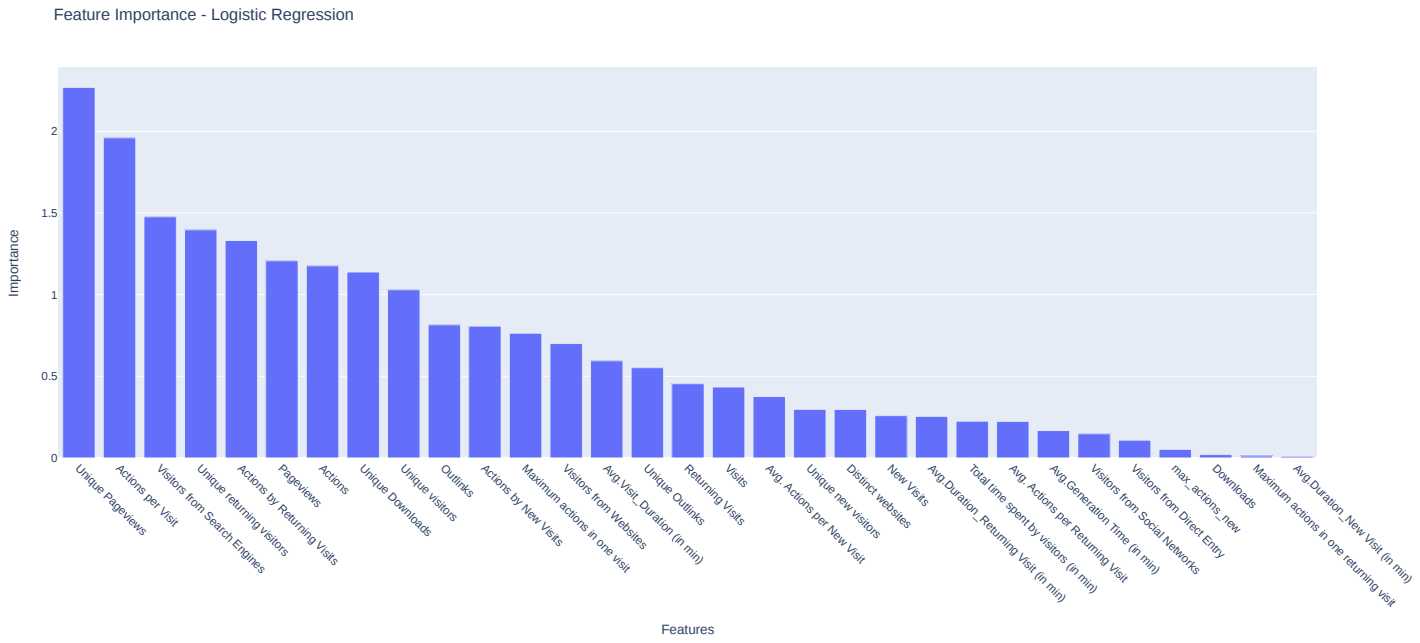
Είναι σημαντικό να ληφθεί υπόψη πως αυτή η αυτοματοποιημένη διαδικασία αφαίρεσης μεταβλητών, που βασίζεται στον τρόπο που διαχειρίζεται στατιστικά κάθε μοντέλο τη σημαντικότητά τους, μπορεί να επηρεάζεται από το μικρό πλήθος δεδομένων και από τις υψηλές συσχετίσεις των μεταβλητών που εντοπίστηκαν σε προηγούμενη παράγραφο και την γενικότερη πολυπλοκότητα του πίνακα. Άλλες μέθοδοι επιλογής ή αφαίρεσης μεταβλητών, περισσότερες πληροφορίες σχετικά με την συγκεκριμένη χρήση αυτών ή σύγκριση με άλλα σύνολα προερχόμενα από Websites, ίσως μπορούν να δώσουν περισσότερες πληροφορίες.

Αν και σε αυτή την περίπτωση η RFE δεν φάνηκε να βελτιστοποιεί τα μοντέλα, προσφέρει μια πολύ καλή εικόνα της ερμηνείας των επιλεχθέντων μεταβλητών και του τρόπου επίδρασης σε κάθε μοντέλο, όπως αναλύεται στη συνέχεια.

5.2.3 Feature Importance και Feature Ranking

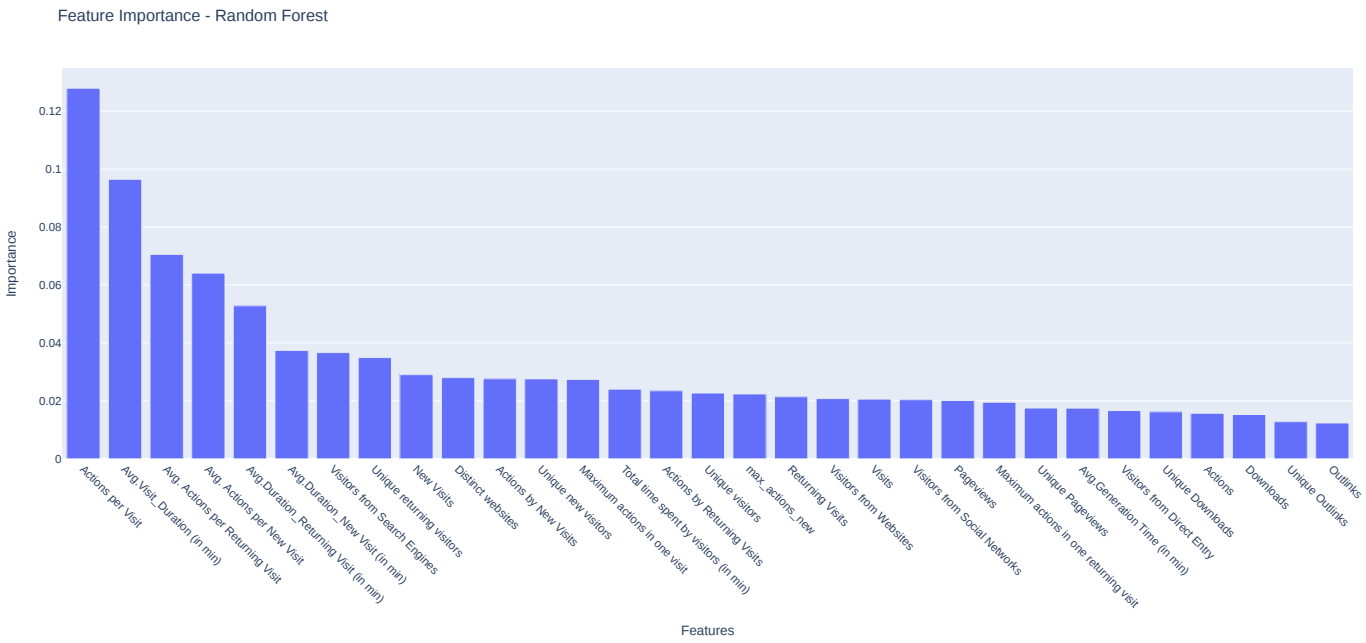
Για την ουσιαστικότερη κατανόηση των μοντέλων και ερμηνείας αποτελεσμάτων στα γενικότερα πλαίσια του XAI. Αρχικά δίνονται τα feature ranking για τη Baseline εφαρμογή και στη συνέχεια για την εφαρμογή με RFE.

5.2.3.1 Feature Importance Baseline



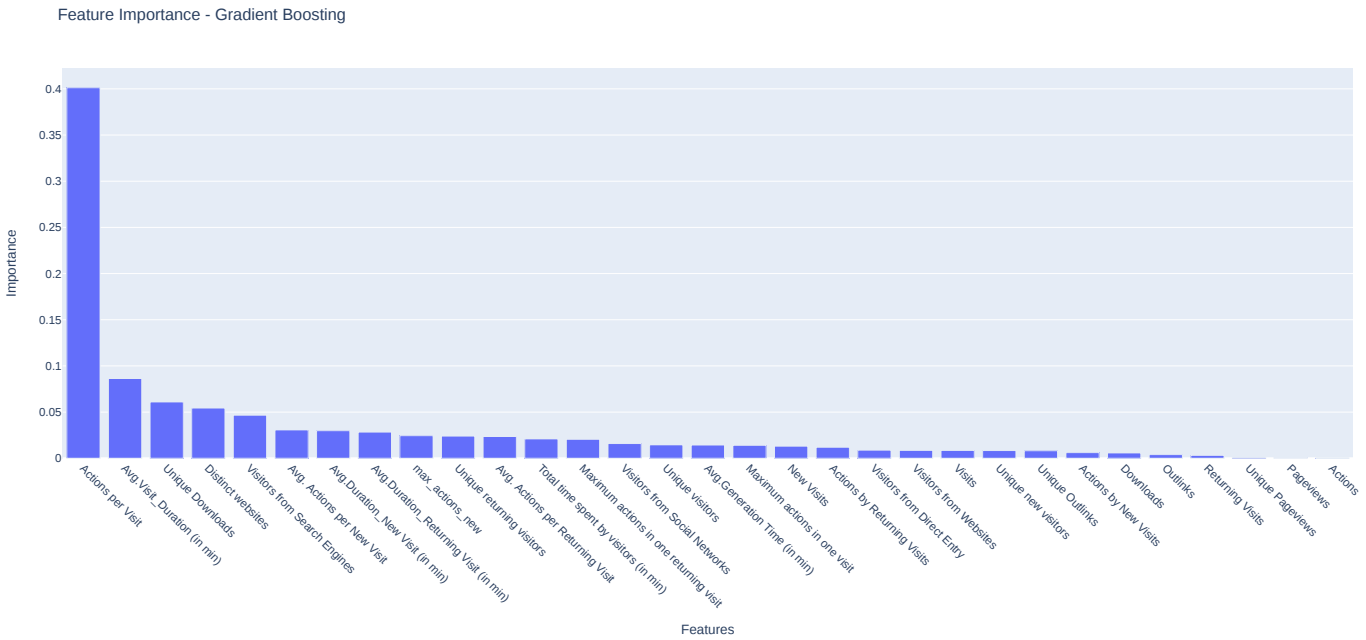
Σχήμα 5.30: Σημαντικότητα μεταβλητών για Logistic Regression

- υψηλότερη συνεισφορά: Unique pageviews, Actions per Visit, Visitors from Search Engines
- χαμηλότερη συνεισφορά: Avg.Duration_New Visit (in min), Maximum actions in one returning visit, Downloads



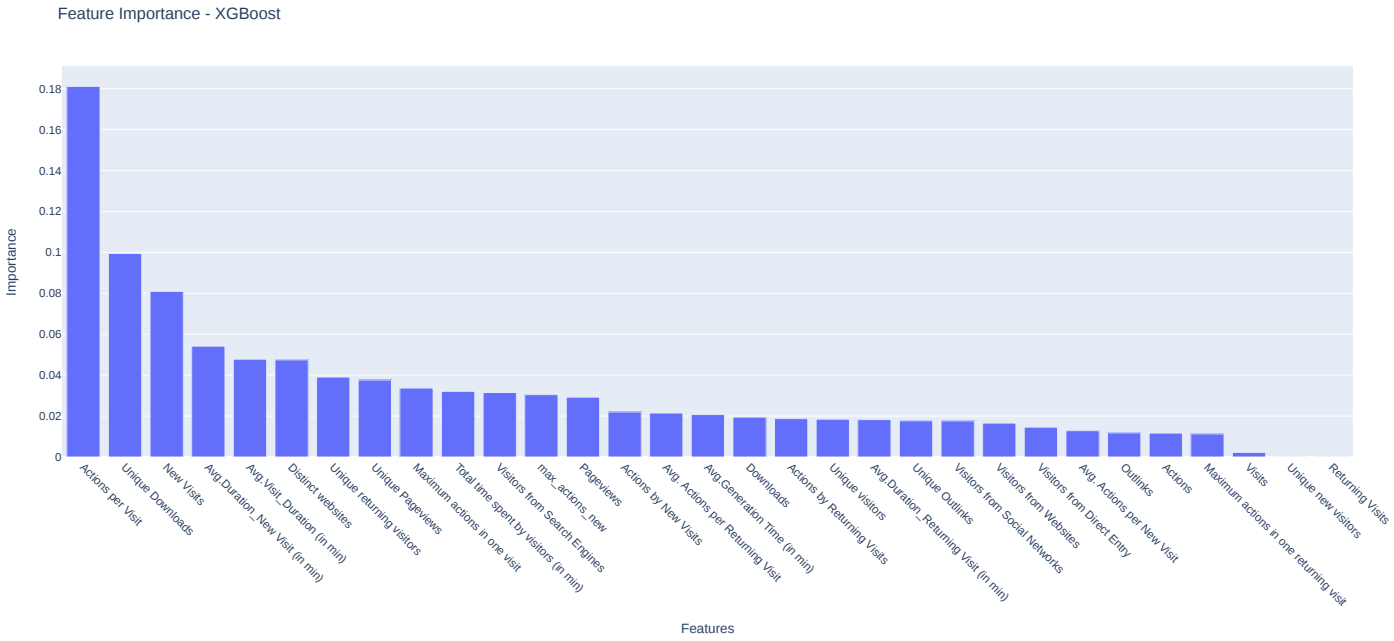
Σχήμα 5.31: Σημαντικότητα μεταβλητών για Random Forest

- υψηλότερη συνεισφορά: Actions per Visit, Avg.Visit_Duration (in min), Avg.Actions per Returning Visit
- χαμηλότερη συνεισφορά: Outlinks, Unique Outlinks, Downloads



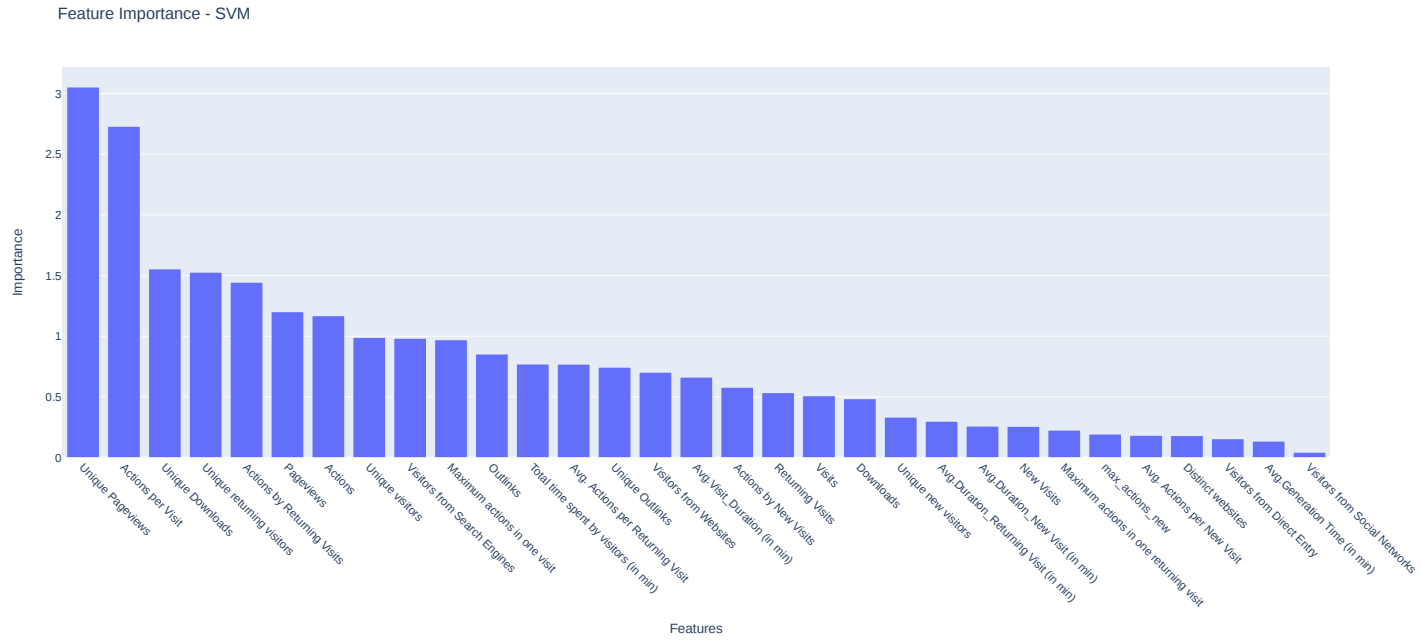
Σχήμα 5.32: Σημαντικότητα μεταβλητών για Gradient Boosting

- υψηλότερη συνεισφορά: Actions per Visit, Avg.Visit_Duration (in min), Unique Downloads
- χαμηλότερη συνεισφορά: Actions, Pageviews, Unique Pageviews



Σχήμα 5.33: Σημαντικότητα μεταβλητών για XGBoost

- υψηλότερη συνεισφορά: Actions per Visit, Unique Downloads, New Visits
- χαμηλότερη συνεισφορά: Returning Visits, Unique new visitors, Visits



Σχήμα 5.34: Σημαντικότητα μεταβλητών για SVM

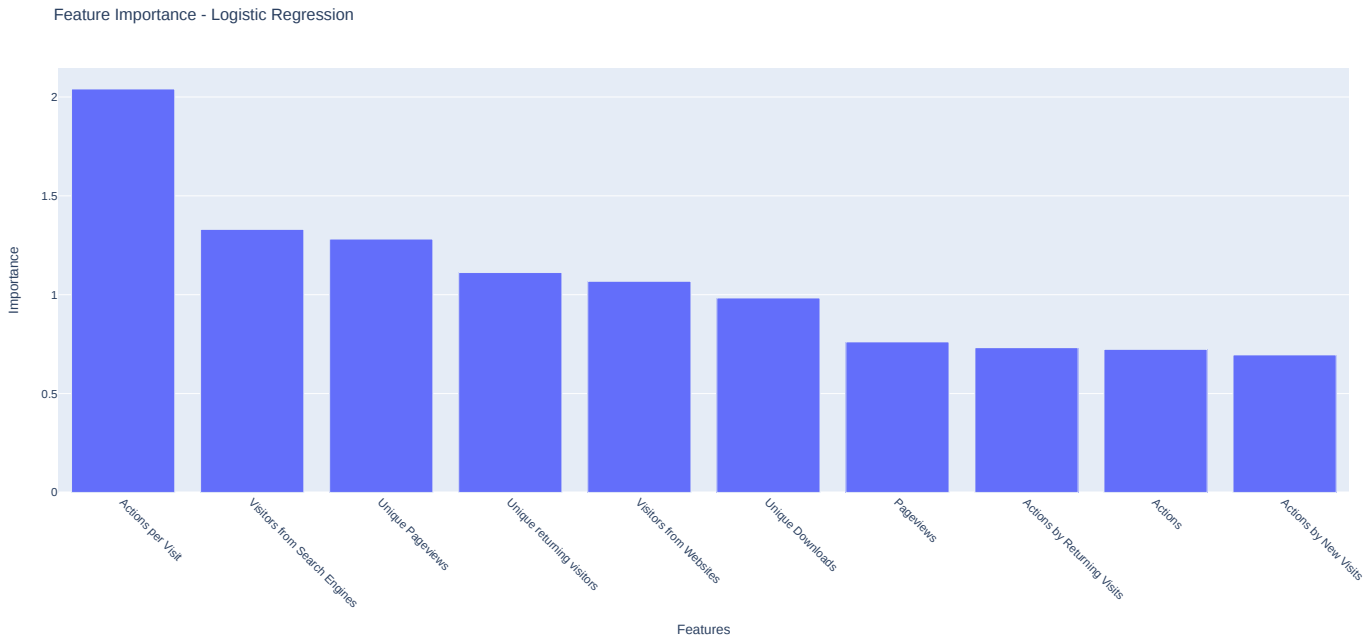
- υψηλότερη συνεισφορά: Unique Pageviews, Actions per Visit, Unique Downloads
- χαμηλότερη συνεισφορά: Visitors from Social Networks, Avg.Generation Tim (in min, Visitors from Direct Entry)

Παρατηρείται πως το Feature Ranking του SVM εμφανίζει αρκετές ομοιότητες με αυτό του Logistic Regression ως προς τις μεταβλητές που συνεισφέρουν περισσότερο, και διαφοροποιούνται στις μεταβλητές με τη λιγότερο σημαντικότητα. Και τα δύο μοντέλα δίνουν προτεραιότητα σε μεταβλητές που δηλώνουν ενέργεια, επισκέπτες που επανασυνδέονται, ή αριθμό σελίδων και μεταφορτώσεων. Αντίθετα, μεταβλητές που δηλώνουν διάρκεια σύνδεσης, είτε συνολικό είτε κατά μέσο όρο επίσκεψης, εμφανίζουν μέτρια ή μικρή σημαντικότητα. Τα δύο αυτά μοντέλα βασίζονται στον υπολογισμό των συντελεστών των μεταβλητών για να κάνουν την κατάταξη, γι΄ αυτό θα πρέπει να ληφθεί υπόψη η φύση των δεδομένων όπως αναλύθηκε σε προηγούμενη παράγραφο που αποτελούνται από ισχυρές συσχετίσεις και πολυσυγγραμμικότητα.

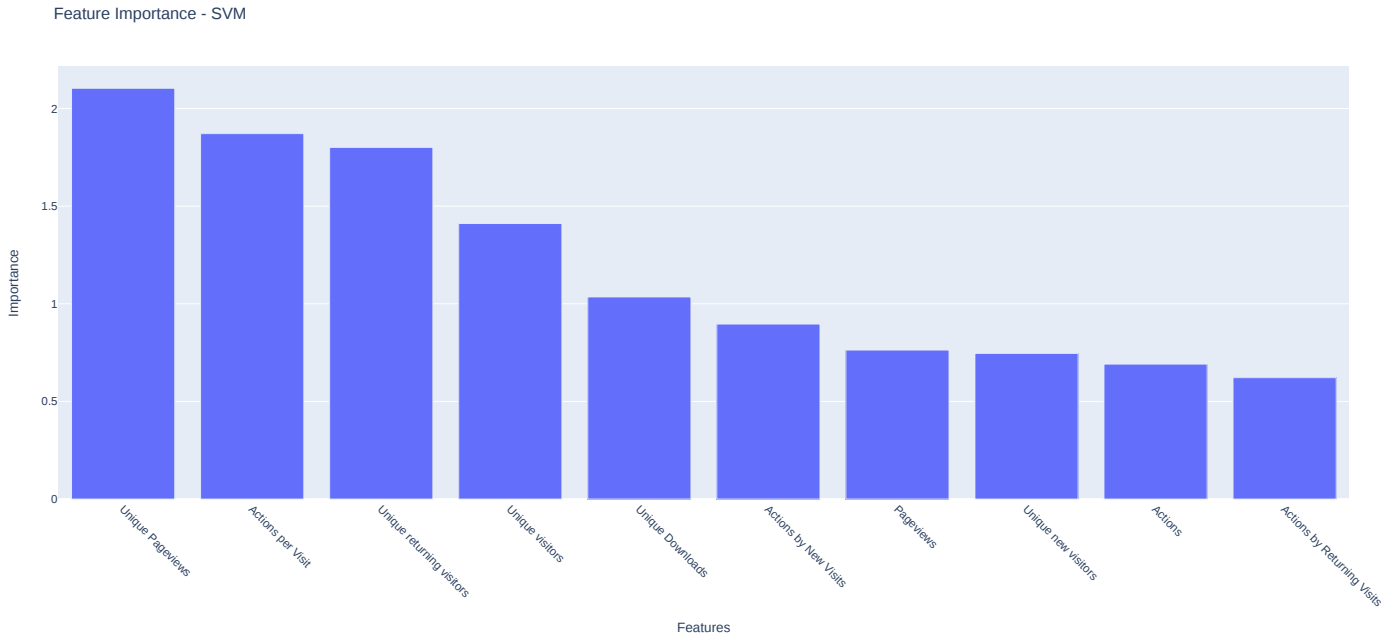
Για τα μοντέλα που βασίζονται σε δέντρα, φαίνεται μια ποικιλομορφία στον τρόπο που κατέταξαν τις μεταβλητές, αν και όλα έθεσαν πολύ υψηλά τα Actions per Visit ως σημαντική μεταβλητή. Συγκεκριμένα για το Random Forest, το αποτέλεσμα φαίνεται να επηρεάζεται περισσότερο, μετά από την Actions per Visit, από χρονικές μεταβλητές μέσου χρόνου παραμονής για τους διαφορετικούς τύπους επισκεπτών (visit, new, returning) καθώς και από τον μέσο όρο ενεργειών αυτών.

Ενδιαφέρον παρουσιάζει η μετρική Unique Downloads που είναι δεύτερο σε σημαντικότητα για το Gradient Boosting και το XGBoost και τρίτο για το SVM, ενώ για το Random Forest είναι προς το τέλος της κατάταξης και για το Logistic Regression στις σχετικές υψηλές θέσεις επιρροής.Να σημειωθεί πως η μετρική Downloads καθώς και Unique Downloads ήταν από τις μετρικές με τις μικρότερες τιμές, μηδενικές σε αρκετές ημερομηνίες. Επομένως ο ρόλος της θα πρέπει να εξεταστεί με προσοχή.

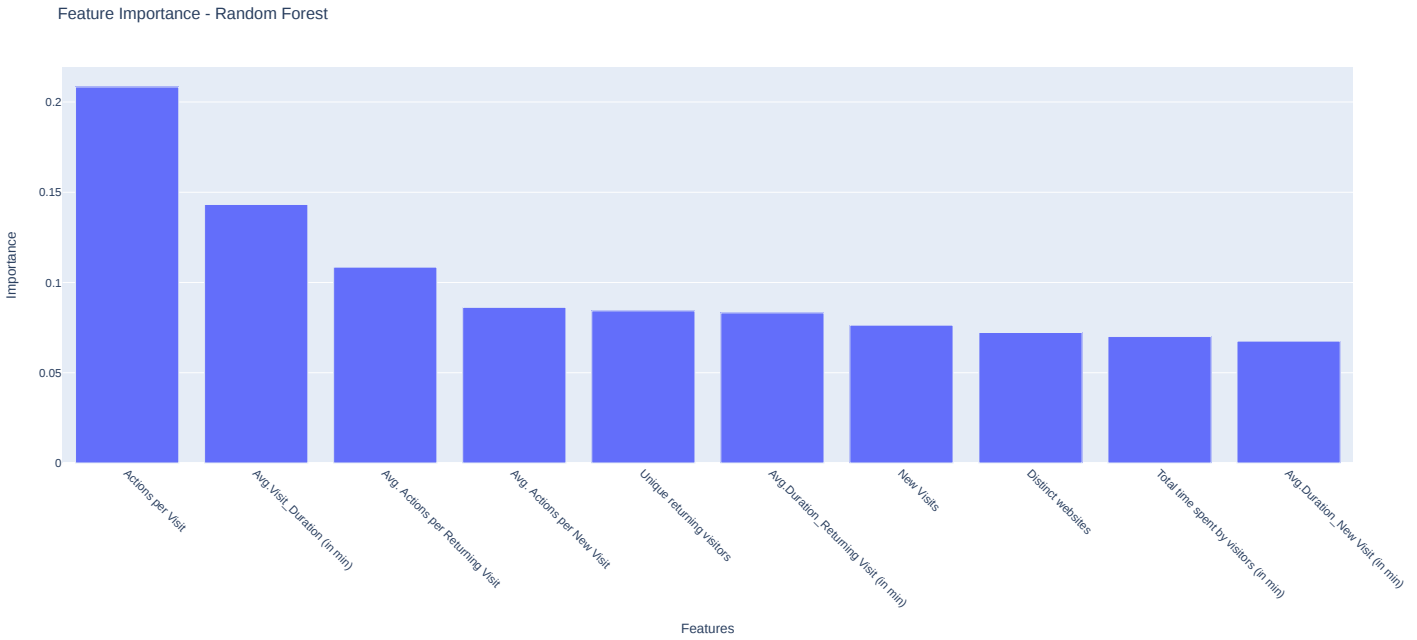
5.2.3.2 Feature Importance με RFE



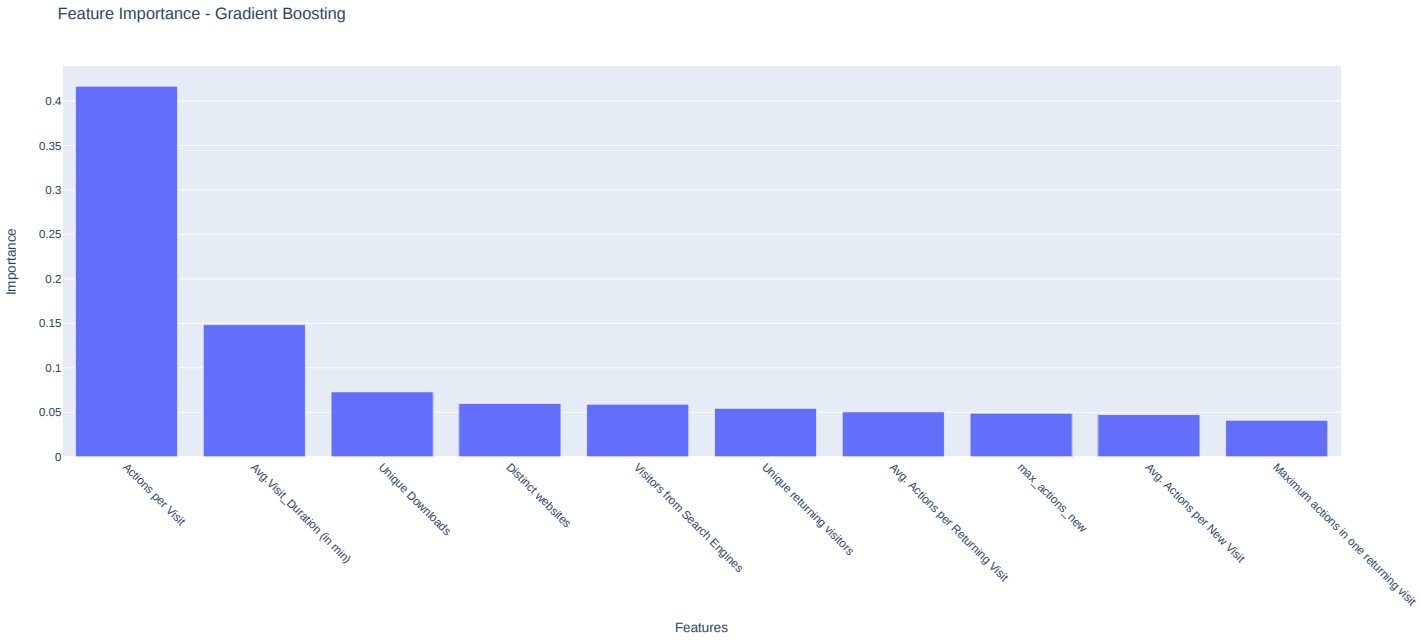
Σχήμα 5.35: Feature Ranking για Logistic Regression μετά την εφαρμογή RFE



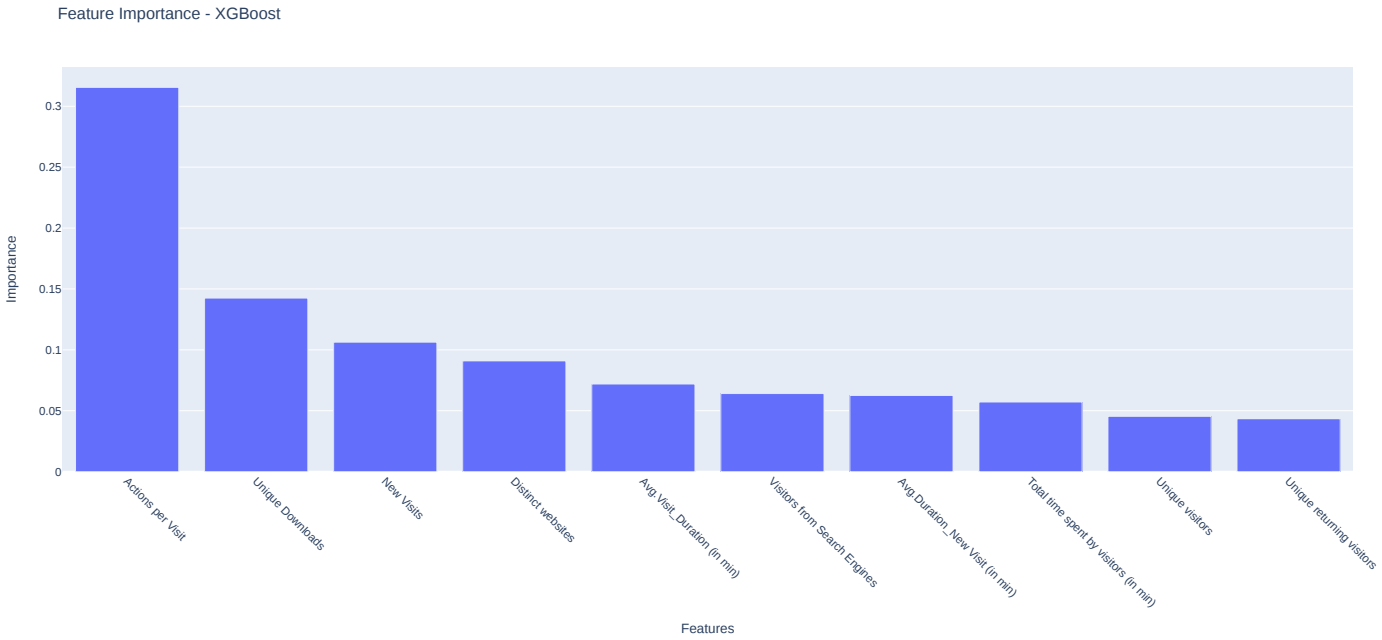
Σχήμα 5.36: Feature Ranking για SVM μετά την εφαρμογή RFE



Σχήμα 5.37: Feature Ranking για Random Forest μετά την εφαρμογή RFE



Σχήμα 5.38: Feature Ranking για Gradient Boosting μετά την εφαρμογή RFE



Σχήμα 5.39: Feature Ranking για XGBoost μετά την εφαρμογή RFE

- Η εφαρμογή της RFE ανέδειξε την Actions per Visit ως σημαντικότερη μεταβλητή σε όλα τα μοντέλα εκτός του SVM που κατατάσσει πρώτη τη Unique Pageviews.

- Logistic Regression και SVM δίνουν προτεραιότητα στις ίδιες μεταβλητές, χωρίς ωστόσο να έχουν την ίδια κατάταξη. Αυτές έχουν να κάνουν κυρίως με τον τύπο επισκεπτών, τον αριθμό ενεργειών και τον αριθμό ενεργειών ανά τύπο επισκεπτών. Εξαίρεση αποτελούν οι μεταβλητές Visitors from Search Engines που το μοντέλο Logistic Regression κατατάσσει δεύτερη και η Visitors from Websites που δεν υπάρχει στην κατάταξη του SVM.

- το Random Forest φαίνεται να προσπαθεί να κάνει ισορροπημένη κατάταξη μεταβλητών αφού εμφανίζει μικρότερες διαφορές της πρώτης σημαντικής (Actions per Visit) από τις υπόλοιπες που ακολουθούν. Το μοντέλο δίνει έμφαση στο μέσο όρο ενεργειών ανά τύπο επισκεπτών (New Visits, Returning Visits) αλλά και στις χρονικές μεταβλητές διάρκειας παραμονής στην πλατφόρμα. Στην κατάταξη εμφανίζονται επίσης οι μεταβλητές New Visits, όπως και στο XGBoost και η Distinct Websites που εμφανίζεται και στα τρία ensemble μοντέλα.

- το XGBoost επηρεάζεται και αυτό από χρονικές μεταβλητές αλλά και από τον τύπο επισκεπτών (New Visits, Unique Visitors, Returning Visits). Ως δεύτερη σημαντικότερη μεταβλητή θέτει την Unique Downloads η οποία εμφανίζεται και στα υπόλοιπα μοντέλα εκτός του Random Forest

- το Gradient Boosting δίνει υψηλή προτεραιότητα στην Actions per Visit κι οι υπόλοιπες ακολουθούν με σημαντική διαφορά. Ως δεύτερη σημαντική κατατάσσει τη χρονική Avg. Visit Duration (in min).

Από τα παραπάνω προκύπτει πως η ταξινόμηση του Bounce Rate(%) επηρεάζεται σημαντικά από τις ενέργειες των επισκεπτών, είτε αυτές που αφορούν το σύνολο των επισκεπτών είτε των διαφόρων τύπων αυτών. Ακόμη, τα μοντέλα Random Forest και XGBoost επηρεάζονται επίσης από χρονικές μεταβλητές, σε λιγότερο ή περισσότερο βαθμό το καθένα.

Εκτός από τις ομοιότητες που εμφανίζουν ορισμένα από τα μοντέλα μεταξύ τους ως προς την κατάταξη των μεταβλητών αξίζει να σημειωθεί και ο διαφορετικός τρόπος λειτουργίας τους ανάλογα τη φύση του αλγορίθμου τους. Τα μοντέλα Logistic Regression, SVM εμφανίζουν αρκετές ομοιότητες στις μεταβλητές που έχουν κατατάξει και ταυτόχρονα διαφέρουν από τα ensemble μοντέλα. Για παράδειγμα, τα γραμμικά μοντέλα δίνουν έμφαση στις ενέργειες ανά τύπο επισκέπτη ενώ τα ensemble, στον μέσο όρο ενεργειών και ενεργειών ανά τύπο επισκέπτη. Παράλληλα διάφορες μετρικές, όπως οι χρονικές, φαίνεται να επηρεάζουν ορισμένα από τα μοντέλα, ωστόσο μία συγχώνευση αυτών ή δημιουργία νέας εξ αυτών, θα μπορούσε να δώσει διαφορετικά και πιο συγκεκριμένα αποτελέσματα που θα βοηθούσαν στη βελτίωση συγκεκριμένων υπηρεσιών, που αφορούν τον χρόνο παραμονής παραδείγματος χάριν.

Για τη βελτίωση των υπηρεσιών μιας πλατφόρμας όπως του Up2U, οι γνώσεις των παραπάνω μπορούν να βοηθήσουν τόσο στο κομμάτι της ανάλυσης όσο και στο κομμάτι προσβασιμότητας της πλατφόρμας και την επίτευξη επιθυμητών τιμών Bounce Rate(%) ή άλλων μετρικών.

Κεφάλαιο 6

Επίλογος

6.1 Σύνοψη

Στα πλαίσια της διπλωματικής έγινε η ανάλυση Web Analytics του Πολωνικού e-Service του προγράμματος Up2U για την περίοδο 8 Νοεμβρίου 2019 - 25 Μαρτίου 2021. Καθώς το διάστημα αυτό περιλαμβάνει την περίοδο της πανδημίας Covid-19 και της υποχρεωτικής τηλεκπαίδευσης, αναλύθηκαν ξεχωριστά τα διαστήματα βάσει των κυμάτων της πανδημίας. Από την ανάλυση προέκυψε πως ενώ η πλατφόρμα είχε περιορισμένη χρήση (πιλοτική εφαρμογή) αρχικά, το πρώτο πανδημικό κύμα μεγιστοποίησε τη χρήση της και έδειξε συστηματική επισκεψιμότητα. Κατά τη διάρκεια του δεύτερου κύματος υπήρξε κλιμακωτή αύξηση χρήσης που μειώθηκε καθώς επανήλθαν τα δια ζώσης μαθήματα. Η διάρκεια παραμονής επισκεπτών που καταγράφηκε στο διάστημα αυτό, η αναλογία ενεργειών και σελίδων που ανοίχτηκαν θεωρούνται σχετικά μικρές δείχνοντας πως οι χρήστες είχαν συγκεκριμένο σκοπό εισαγωγής, ή και πως η πληροφορία που αναζητούσαν ήταν προσβάσιμη και άμεσα διαθέσιμη χωρίς να απαιτούνται περαιτέρω ενέργειες. Ενδιαφέρον έχει η αλλαγή συμπεριφοράς των επισκεπτών ως προς τα κανάλια εισόδου. Αρχικά, βασική πηγή πρόσβασης ήταν η απευθείας σύνδεση, ενώ κατά τη διάρκεια του πρώτου κύματος είσοδος από άλλες σελίδες αποτέλεσε βασικό κανάλι. Κατά την τελευταία περίοδο παρατηρήθηκε σημαντική αύξηση στην είσοδο από μηχανές αναζήτησης είτε λόγω προωθητικών ενεργειών, είτε εξοικείωσης με την πλατφόρμα.

Το δεύτερο μέρος της διπλωματικής αφορούσε την εφαρμογή μοντέλων μηχανικής μάθησης για ταξινόμηση της μεταβλητής Bounce Rate(%) και ερμηνεία των μεταβλητών που την επηρεάζουν. Τα μοντέλα εφαρμόστηκαν στην Baseline μορφή τους, στην απλή δηλαδή μορφή τους με τις προεπιλεγμένες παραμέτρους τους, με και χωρίς τη μέθοδο SMOTE, και στη συνέχεια με τη μέθοδο RFE. Η μέθοδος SMOTE δεν φάνηκε να ευνοεί σε αξιοσημείωτο βαθμό τα αποτελέσματα και για αυτό απορρίπτεται καθώς στα δεδομένα υπάρχει ήδη θόρυβος. Σε όλες τις περιπτώσεις παρατηρήθηκε έντονο overfitting, κυρίως για μοντέλα ensemble, ενώ τα γραμμικά μοντέλα Logistic Regression, SVM είχαν καλύτερη προσαρμογή αλλά χαμηλότερες αποδόσεις. Random Forest, και XGBoost είχαν τις καλύτερες αποδόσεις σε όλες τις περιπτώσεις ενώ το Random Forest μπόρεσε να κάνει καλύτερες ταξινομήσεις. Μετά την εφαρμογή RFE η ακρίβεια όλων των μοντέλων μειώθηκε, παρόλα αυτά δίνεται η ευκαιρία για καλύτερη ερμηνεία των μεταβλητών. Να σημειωθεί πως στα δεδομένα υπήρχε έντονο

πρόβλημα πολυσυγγραμμικότητας, το πλήθος των δεδομένων ήταν αρκετά μικρό (504 παρατηρήσεις), πράγμα που επηρέασε τις αποδόσεις αλλά και τους τρόπους που ανταποκρίθηκαν τα μοντέλα, για παράδειγμα τα γραμμικά μοντέλα φάνηκε να επηρεάζονται σε μεγαλύτερο βαθμό από τις διάφορες εφαρμογές.

Ως προς την ερμηνεία των μεταβλητών, υπήρξε ομοιότητα στον τρόπο που κατέταξαν τις μεταβλητές τα γραμμικά μοντέλα και αντίστοιχα ομοιότητες στα μοντέλα ensemble. Τα γραμμικά μοντέλα έδωσαν προτεραιότητα σε μεταβλητές Actions (π.χ. Actions, Actions by New Visits, Actions per Visits κ.λ.π.), καθώς και στο είδος επισκέπτη, ενώ τα μοντέλα ensemble φάνηκε να επηρεάζονται και από χρονικές μεταβλητές (π.χ. Avg. Visit Duration, Total time spent by visitors). Υψηλότερη σε σημαντικότητα κατατάχθηκε η Actions per visit, εκτός από το μοντέλο SVM το οποίο επιλέγει ως σημαντικότερη την Unique Pageviews.

6.2 Προοπτικές Μελλοντικής Επεκτάσεων

Το συγκεκριμένο σύνολο δεδομένων με την πολυπλοκότητα που εμφάνισε και τα αποτελέσματα που προέκυψαν ενδείκνυται για περαιτέρω μελέτη και ερμηνεία. Μία προσέγγιση που θα μπορούσε να ακολουθηθεί, βάσει και της πορείας της έρευνας που προηγήθηκε, είναι μία πιο ενδελεχής κατανόηση και ερμηνεία των αποτελεσμάτων των μοντέλων που εφαρμόστηκαν. Το πεδίο eXplainable Artificial Intelligence προσφέρει πληθώρα μοντέλων τα οποία θα μπορούσαν να επεκτείνουν την διάγνωση και ερμηνεία των μοντέλων που ξεκίνησε από αυτή τη διπλωματική. Τέτοια μοντέλα, όπως το SHAP (SHapley Additive exPlanations), σε συνδυασμό με τους στόχους που προσπαθεί να πετύχει ένα πρόγραμμα όπως το Up2U μπορούν να οδηγήσουν σε ορθότερες λήψεις αποφάσεων σχετικά με την προσέγγιση νέων χρηστών, εφαρμογής ενεργειών σε μία πλατφόρμα ή άλλων σημείων ενδιαφέροντος που προκύπτουν από τη μελέτη και ερμηνεία των μεταβλητών.

Ενδιαφέρον θα είχε ακόμη, η εφαρμογή και μεθοδολογία που ακολουθήθηκε για τα δεδομένα του Πολωνικού e-Service να εφαρμοστούν και για κάποιο άλλο e-Service του Up2U ή να συγκριθούν με άλλες πλατφόρμες εξολοκλήρου ώστε να διερευνηθεί η μετρική Bounce Rate(%) σε ένα πιο γενικό πλαίσιο. Μέθοδοι όπως υπερπαραμετροποιήσεις των μοντέλων, διαφορετικές επιλογές μεταβλητών ή συνδυασμός αυτών θα μπορούσε να αποτελεί έναν άλλο τρόπο προσέγγισης για τη βελτίωση της απόδοσης των μοντέλων.

Bibliography

[1] I. Bekavac and D. Garbin Praničević, «Web analytics tools and web metrics tools: An overview and comparative analysis», *Croatian Operational Research Review*, vol. 6, no. 2, pp. 373–386, 2015.

[2] J. Zheng and S. Peltsverger, «Web analytics overview», in Jan. 2015.

[3] U. Jyothi, S. Bonthu, and B. V. Prasanthi, «A study on raise of web analytics and its benefits», *International Journal of Computer Sciences and Engineering*, vol. 5, no. 10, pp. 59–64, Oct. 2017.

[4] T. Rohloff, S. Oldag, J. Renz, and C. Meinel, «Utilizing web analytics in the context of learning analytics for large-scale online learning», *2019 IEEE Global Engineering Education Conference (EDUCON)*, pp. 296–305, Apr. 2019.

[5] V. Kumar and G. A. Ogunmola, «Web analytics for knowledge creation», *International Journal of Cyber Behavior, Psychology and Learning*, vol. 10, no. 1, pp. 1–14, Winter 2020.

[6] Pufferr, *A machine learning study of the google analytics metrics predicting content quality*, Accessed: 2023-10-07, 2023. [Online]. Available: https://pufferr.co.uk/a-machine-learning-study-of-the-google-analytics-metrics-predicting-content-quality/.

[7] *Bounce rate benchmarks: What's a good bounce rate, anyway?*, https://cxl.com/guides/bounce-rate/benchmarks/, Accessed: 2023-10-07, 2023.

[8] *The ultimate guide to bounce rate*, https://cxl.com/guides/bounce-rate/, Accessed: 2023-10-07, 2023.

[9] J. Pomerantz, M. Brown, and D. C. Brooks, «Foundations for a next generation digital learning environment: Faculty, students, and the lms», ECAR, Louisville, CO, Research report, Jan. 2018, ©2018 EDUCAUSE. Licensed under a Creative Commons BY-NC-ND 4.0 International License. [Online]. Available: https://library.educause.edu/resources/2018/1/foundations-for-a-next-generation-digital-learning-environment-faculty-students-and-the-lms.

[10] E. Kikkenborg, C. Dreef, M. Grammatikou, *et al.*, «Deliverable d1.2 final project report, ipr and external liaison summary», GÉANT Association, Work Package 1, Tasks 1.1, 1.2, 1.3, 2020, Horizon 2020 research and innovation programme (Grant Agreement No. 732049). [Online]. Available: https://owncloud.gwdg.de/index.php/s/H45wJAI9EOCgiM3.

[11] B. Tóth, N. Dias, and P. Szegedi, «Deliverable d2.2 dissemination and outreach report year 1», Kormányzati Informatikai Fejlesztési Ügynökség (KIFÜ), Work Package 2, Task 2.1, Oct. 2018, Horizon 2020 research and innovation programme (Grant Agreement No. 732049). [Online]. Available: https://owncloud.gwdg.de/index.php/s/pWG08J6VbuDFfdH.

[12] I. Barth, E. Bertazzon, G. Cibulskis, *et al.*, «Deliverable d2.3 dissemination and outreach report year 2», GÉANT Association, Work Package 2, Task 2.1, Jan. 2019, Horizon 2020 research and innovation programme (Grant Agreement No. 732049). [Online]. Available: https://owncloud.gwdg.de/index.php/s/PutItsvPhaNr5mA.

[13] A. Third, D. Pantazatos, A. Rodríguez Pereira, *et al.*, «Deliverable d4.1 application toolbox design and prototype», Open University (OU), Work Package 4, Tasks T4.1, T4.2, T4.3, Jun. 2017, Horizon 2020 research and innovation programme (Grant Agreement No. 732049). [Online]. Available: https://owncloud.gwdg.de/index.php/s/CZRzol7dkYbEGal.

[14] M. Grammatikou, A. Third, D. Pantazatos, *et al.*, «Deliverable d4.2 enhanced application toolbox and technical specifications», National Technical University of Athens (NTUA), Work Package 4, Tasks T4.1, T4.2, T4.3, Jun. 2018, Horizon 2020 research and innovation programme (Grant Agreement No. 732049). [Online]. Available: https://owncloud.gwdg.de/index.php/s/PYwe36YWYyBPsYi.

[15] M. Zimniewicz, Z. Karwasiński, T. Kuczyński, *et al.*, «Deliverable d7.4 final demonstration and report on large-scale pilots», Poznań Supercomputing and Networking Center (PSNC), Work Package 7, Tasks 1, 2, 3, 4, Jun. 2020, Horizon 2020 research and innovation programme (Grant Agreement No. 732049). [Online]. Available: https://owncloud.gwdg.de/index.php/s/cJ3uVwzfNGpwQPZ.

[16] B. Patil, R. Joshi, and D. Toshniwal, «Missing value imputation based on k-mean clustering with weighted distance», vol. 94, Aug. 2010, pp. 600–609, ISBN: 978-3-642-14833-0.

[17] P. D. Allison *et al.*, *Missing data*. Sage Thousand Oaks, CA, 2010, vol. 200210.

[18] J. Kaiser, «Dealing with missing values in data.», *Journal Of Systems Integration (1804-2724)*, vol. 5, no. 1, 2014.

[19] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[20] T. Cleff, «Exploratory data analysis in business and economics», *Exploratory Data Analysis in Business and Economics. https://doi. org/10.1007/978-3-319-01517-0*, 2014.

[21] G. J. Myatt, *Making sense of data: a practical guide to exploratory data analysis and data mining*. John Wiley & Sons, 2007.

[22] G. J. Myatt and W. P. Johnson, *Making sense of data II: A practical guide to data visualization, advanced data mining methods, and applications*. John Wiley & Sons, 2009, vol. 2.

[23] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.

[24] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: a Guide for Data Scientists*. " O'Reilly Media, Inc.", 2016.

[25] A. Ullah, U. Qamar, F. H. Khan, and S. Bashir, «Dimensionality reduction approaches and evolving challenges in high dimensional data», in *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, 2017, pp. 1–8.

[26] J. Dunn, L. Mingardi, and Y. D. Zhuo, «Comparing interpretability and explainability for feature selection», 2021.

[27] Ε. Κ*ύρκος*, «Ε*πιχειρηματική ευφυΐα και εξόρυξη δεδομένων*», 2016.

[28] Ο*ικονόμου* and Χ. Κ*αρώνη*, «Σ*τατιστικά* Μ*οντέλα* Π*αλινδρόμησης*», *Εκδόσεις Συμεών Αθήνα*, 2010.

[29] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[30] L. Breiman, «Random forests», *Machine learning*, vol. 45, pp. 5–32, 2001.

[31] P. Misra and A. Yadav, «Improving the classification accuracy using recursive feature elimination with cross-validation», vol. 11, pp. 659–665, May 2020.

[32] G. James, D. Witten, T. Hastie, R. Tibshirani, *et al.*, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[33] S. S. Skiena, *The data science design manual*. Springer, 2017.

[34] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, «Expanding explainability: Towards social transparency in ai systems», in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.

[35] I. Guyon and A. Elisseeff, «An introduction to variable and feature selection», *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.